

Regressionsanalyse mit SPSS

SPSS回归分析

[德] Dr. Christian FG Schendera (克里斯蒂安·FG·申德拉博士) 著

宋武 译

李洪成 审校

電子工業出版社

Publishing House of Electronics Industry

北京•BEIJING

内 容 简 介

回归分析在科学研究领域是最常用的统计方法。本书介绍了一些基本的统计方法,例如,相关、回归(线性、多重、非线性)、逻辑(二项、多项)、有序回归和生存分析(寿命表法、Kaplan-Meier 法以及 Cox 回归)。后面的章节介绍了另外一些回归分析方法和模型,例如,个体生长曲线的建模、PLS 部分最小平方回归、岭回归、巢式病例对照研究。

本书对运用 SPSS 进行回归分析的介绍,目的是让读者对于这方面的基础知识有一个初步了解和掌握,有经验的读者藉此可在数据挖掘(例如,利用 Clementine)领域独立地继续学习新知识。

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有,侵权必究。

图书在版编目(CIP)数据

SPSS 回归分析 / (德) 申德拉 (Schendera, C.F.G.) 著; 宋武译. —北京: 电子工业出版社, 2015.3
ISBN 978-7-121-22300-6

I. ①S… II. ①申… ②宋… III. ①统计分析—回归分析—软件包 IV. ①C819-39

中国版本图书馆 CIP 数据核字 (2014) 第 002788 号

策划编辑: 张月萍

责任编辑: 徐津平

印 刷: 北京天宇星印刷厂

装 订: 北京天宇星印刷厂

出版发行: 电子工业出版社

北京市海淀区万寿路 173 信箱

邮编: 100036

开 本: 787×1092 1/16

印张: 21.25 字数: 558 千字

版 次: 2015 年 3 月第 1 版

印 次: 2015 年 3 月第 1 次印刷

印 数: 3000 册

定价: 78.00 元

凡所购买电子工业出版社图书有缺损问题, 请向购买书店调换。若书店售缺, 请与本社发行部联系, 联系及邮购电话: (010) 88254888。

质量投诉请发邮件至 zltz@phei.com.cn, 盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线: (010) 88258888。

前言

回归分析及其各种变型在科学和研究领域是最常用的统计方法（例如，参见著作 Hsu, 2005、Pötschke & Simonson, 2003、Elmore & Woehlke, 1998, 1996 和 Goodwin & Goodwin, 1985）。各个学科领域、科研工作以及统计学的发展既对这些统计方法不断提出更高的要求，同时又起到巨大的推动作用（例如，参见 Rigby 等人 2004 年的著作和 Ripoll 等人 1996 年的著作）。

某些作者也将回归分析称为最古老的统计方法之一。Stanton (2001) 认为，线性回归基本统计方法的建立起源于 Karl Pearson（例如，1896 年发表的论文）的论文。Howarth (2001) 则认为，线性分析基本理论及其使用方法甚至要追溯到 Bond 在 1636 年发表的一篇论文（例如，参见 Finney, 1996）。

回归分析发展史上的里程碑，是 18 世纪中叶人们创建了基本的计算方法后，从而与此相关地发明了“最小二乘”算法。直到 20 世纪中期发明了计算机之后，多元回归分析方法才被越来越多地应用于海量（并且容易出错）计算，从而加快了回归分析方法的应用、传播和发展。回归分析方法发展史上的其他（任意选出的）里程碑如有 20 世纪 70 年代出现的岭回归、80 年代兴起的稳健回归以及 90 年代创建的新方法或对原有方法的混合使用。但是，作为一种表面上传统的方法，如今回归分析在用于数据挖掘时不仅仅是与其他新型方法（如神经网络）处于同一水平（例如，SPSS, 2007b, 第 10 章、Rud, 2001、Berry & Linoff, 2000 和 Graber, 2000），而且比其他方法的使用要频繁得多（例如，Rexer 等人 2007 年的著作和 Ayres, 2007）。但是，即使是应用数据挖掘的操作十分容易，数据挖掘也不能取代统计学或者计算机信息学知识，而是以这些知识为前提（Schendera, 2007、Khabaza, 2005、Chapman 等人，1999）。因此，本书针对如何应用 SPSS 回归分析方法而介绍的基础知识也可以使读者对数据挖掘领域初窥门径。

统计学的初学者可能对“回归”方法这个大家族的强大、多样化和灵活性感到十分吃惊。对于进阶学习者而言，他们可能始终感兴趣的是如何利用回归分析方法来处理大量的、参差不齐的问题。例如，简单和多重（非）线性回归分析、个体生长曲线、生存分析、时间序列分析等。从另一方面来看，如此广阔的应用范围就面临一个问题，即人们无法单纯依靠背诵就可以

掌握：“对于带有一个定距因变量的线性因果模型，人们通常使用线性回归；对于带有一个二元因变量的因果模型，就使用逻辑回归等。”或者：“对于线性回归使用 SPSS 过程命令 REGRESSION、对于逻辑回归使用 SPSS 过程命令 LOGISTIC 或 NOMREG、对于生存数据使用 SURVIVAL、KM 或 COXREG 等。”

与之相反，包括 SPSS 在内的统计学知识是十分复杂、灵活和多样化的。高级应用者主要将 SPSS 过程命令 REGRESSION 用于线性、共线性数据或者时间序列数据。例如，SPSS 过程命令 GLM 既可以用于在一般线性模型（ALM）中对方差分析的计算，也可以利用多个因变量进行回归分析。例如，相对比较新的 SPSS 过程命令 GENLIN（SPSS 15 版之后才有），作为一般线性模型既可以对伽玛回归、采用重复测量设计的二元逻辑回归进行计算，也可以对区间截尾的生存数据的双对数回归进行计算。

此外，统计量的多样性比 SPSS 本身的功能范围要大得多。例如，生存数据可以分为一个预期的风险（常规的生存分析、生存分析）、多个预期的互斥风险生存分析（competing risk survival analysis）或者反复性风险（recurrent risk survival analysis）。

因此，选择适当的回归分析方法不仅取决于可用的 SPSS 菜单、SPSS 过程命令或者“菜谱”，还取决于具体的内容和方法逻辑。例如，需要调查的问题（假设有很多类型，如区别与关联相比照）。但是也取决于需要确定的定义，例如，数据的测量水平、分布、转换、数据关联性/无关联性、主效应和交互效应的建模及很多其他的定义。应与有经验的方法学专家或者统计学专家协商后，再对回归分析方法做出选择。对于特殊的问题，标准软件也有可能无法实现所需的统计方法。在这种情况下，也可以自己用 SPSS 或者 Python 编程设计出一种方法（参见宏“Ridge-Regression.sps”），或者使用专门的分析软件。进行模型设定和推断性统计假设检验的操作方法通常越来越复杂（参见 Schendera 著作，2007，401-403）。

本书介绍了一些基本的统计方法。例如，相关、回归（线性、多重、非线性）、逻辑（二项、多项）、有序回归和生存分析（寿命表法、Kaplan-Meier 法以及 Cox 回归）。后面的章节介绍了另外一些回归分析方法和模型（例如，个体生长曲线的建模、PLS 部分最小平方回归、岭回归、巢式病例对照研究）。

在这里对 SPSS 进行回归分析的介绍，目的是让读者对于这方面的基础知识有一个初步了解和掌握，有经验的读者可在此基础上在数据挖掘（例如，利用 Clementine）领域独立地继续学习新知识。由于篇幅所限，本书没有介绍很多其他的回归形式和具体应用（例如，非参数回归、分类回归、Weibull 回归、Hedonic 回归等，对此请参见本书第 6 章）。

本书介绍了相关分析（第 1 章）、回归分析（线性、多重、非线性，第 2 章）、逻辑和有序回归分析（第 3 章）以及生存分析（Survival analyse，第 4 章）的基本方法。对于所有的方法，详细阐述了其前提条件和常犯的错误。第 5 章介绍了回归分析的特殊用途（偏回归、个体生长曲线的建模、岭回归）。第 6 章介绍了 SPSS 的其他用途（例如，对多个因变量的回归分析）。

书中的大量计算实例系统地演示了所提出的问题、各个统计量的调用方法（通过鼠标、语法）以及对 SPSS 输出结果的解释。也探讨了各种错误和难点。关于在实施统计分析之前对数据的检验，可参考《SPSS 的数据质量》一书（Schendera，2007）。

书中用单独的段落归纳了实施各个分析的各种前提条件以及对其进行检验的方法。本书浅显易懂，既侧重具体应用，同时对各种方法的解释又没有忽略其复杂性和必要的深度。本书的读者既可以是回归分析的入门者，也可以是经济、生物和社会科学的学生或学者。

本书既采用了菜单导航，又罗列了大量 SPSS 语法。Windows SPSS 的初学者应从中了解到，单击鼠标就可以自动调用 SPSS 语法或者自己编程设计 SPSS 语法（参见 Schendera, 2007, 2005）。针对 SPSS 程序员（也包括普通用户），则展示了如何借助 SPSS 过程命令 PLS 来扩展 SPSS-Python。利用同名的 SPSS 宏介绍了岭回归。

第 1 章介绍了往常被人们低估的相关分析（SPSS 过程命令 CORRELATIONS）的入门知识。本章开头部分解释了关联（因果性），并列举了几个错误结论的例子，如人们常说的喜欢玩暴力型电脑游戏和个人暴力倾向之间的关联。如果读者对回归分析感兴趣，则强烈建议先阅读关于相关分析的章节。借助于相关分析，作者阐述了首要的、对于（线性）回归分析也适用的前提条件。例如，尺度水平、同方差性和连续性。接下来的几段阐述了线性、产生错觉相关和一型误差累积几个主题，并且解释了为什么只给定相关系数的数值是远远不够的。本章还介绍了相关分析的其他一些特殊用途，主要是相关系数的比较和正准相关。最后一段归纳了实施相关分析的各种前提条件以及对其进行检验的方法。

第 2 章介绍了回归分析的入门知识。本章采用逐步推进的架构，来阐明进行回归分析的基本原则，并且帮助读者从一开始就避免常犯的基本错误。

第 2.1 节首先介绍了简单线性回归分析（SPSS 过程命令 REGRESSION）。第 2.1 节是基于第 1 章的内容展开的。通过一个简单的例子，阐述了如何根据杠杆值和残差来检验线性和识别离群值。还阐述了如何检验可能存在的自相关。一般来说，利用线性回归分析只能调查线性函数。利用线性回归分析来调查非线性函数通常会产生错误的结果。

第 2.2 节阐述了如果数据不是线性而是曲线分布时应该怎么做。第 2.2 节是基于第 2.1 节的内容展开的。本章提供了两种解决方案：将非线性函数进行线性化，并用线性回归进行分析；也可以用非线性回归对非线性函数进行估计（SPSS 过程命令 CNLR 和 NLR）。非线性回归是本章的中心主题，包括带有两个预测变量的非线性回归。此外，本节还阐述了用于（非）线性曲线拟合的 SPSS 过程命令 CURVEFIT 的意义和限制。最后几段总结了非线性回归的各种假设，并通过一个总览表介绍了较为知名的一些非线性回归模型，其中含有一个或多个预测变量。

第 2.3 节介绍了多元线性回归分析（SPSS 过程命令 REGRESSION）的基础知识。第 2.3 节是基于第 2.2 节的内容展开的。模型含有多个自变量，而非仅有一个自变量，这种模型主要是在于自变量相互之间的关系。本节着重探讨了建模、变量选择、多重共线性和其他难点。除了识别和消除多重共线性外，本节还探讨了如何处理时间相依（自回归）数据。最后一段归纳了实施（非）线性回归分析的各种前提条件以及对其进行检验的方法（见第 2.4 节）。

第 3 章介绍了逻辑回归和有序回归的基本方法。本章的结构是根据因变量的尺度水平构建的。最后几段分别归纳了所介绍方法的各种前提条件以及对其进行检验的方法。

二元逻辑回归（SPSS 过程命令 LOGISTIC REGRESSION，第 3.2 节）需要使用一个二值因变量，这个方法中没有考虑因变量中的极差信息。本节首先介绍了作为基本方法的二元逻辑回归，然后阐述了这种方法与其他方法的共同点和区别（主要是模型和尺度水平），并根据几

个计算实例，主要阐述变量选择的不同方法，以及对所输出统计量的解释。最后探讨了经常出现的模型拟合优度和预测精度不一致问题。

有序回归（SPSS 过程命令 PLUM，第 3.3 节）需要使用至少两个取值的（定序）因变量，并且考虑到了因变量中的极差信息，同时阐述了与其他方法的共同点和区别（主要是模型、尺度水平）。最后根据几个计算实例，阐述了如何解释模型的 SPSS 输出结果，其中这些模型带有定距和分类预测变量。

多项逻辑回归（SPSS 过程命令 NOMREG，第 3.4 节）同样需要使用一个至少二级的定类因变量，这种方法没有考虑因变量中的极差信息。对多项逻辑回归的阐述与第 3.2 节类似。此外，还介绍了一种特殊情况，即带有定量预测变量的巢式病例对照研究（1:1）。

第 4 章介绍了生存分析的基本方法。原则上，生存分析调查的是到出现特定目标事件为止的时间。目标事件既可以是期望事件（例如，延长订单、受聘、学习成功、治愈等），也可以是不良事件（例如，被解雇、故障、旧病复发、死亡等）。这些方法有各种各样的名称，例如，寿命分析、生存分析、时间影响或者事件分析等，它们来自于对目标事件的不同评估。根据对目标事件的评估结果不同，对于图表应给予不同的解释。

第 4.1 节首先介绍了生存分析的基本原则，然后介绍了面临的一些典型问题和生存分析的目标。

第 4.2 节阐述了对不同生存函数（主要包括累积生存函数 $S(t)$ 、1 减去生存函数 $(1-S(t))$ 、密度函数 $f(t)$ 、对数生存函数 $l(t)$ 以及风险函数 $h(t)$ ）的规定。

第 4.3 节介绍了数据截尾的入门知识。在进行生存分析时，在某些个案中，可能出现目标事件没有如期望的那样发生，也就是说，目标事件完全没有或者没有按期望的（设定的）原因而发生。为了将这些个案与带有期望事件的个案隔开，就需要借助于截尾将其标出。本节介绍了左截尾、右截尾和区间截尾，并且诠释了在（非）试验性调查设计时的截尾。

第 4.4 节以寿命表法和 Kaplan-Meier 法为例，阐述了如何用这些方法测算生存函数，以及在这个过程中如何处理截尾的个案。从第 4.6 节开始介绍 SPSS 示例。

第 4.5 节介绍了对各组进行比较的不同检验：对数极差检验（又称时序检验或者 Mantel-Cox 检验）、Breslow 检验（又称修正的 Wilcoxon 检验、Wilcoxon 秩和检验）、Tarone-Ware 检验和似然比检验。此外，本章还归纳了一个比较性综述，以及对于解释这些检验的建议方案。

在第 4.6 节中，计算和解释了如何用 SPSS 的寿命表法（SPSS 过程命令 SURVIVAL）和 Kaplan-Meier 法（SPSS 过程命令 KM）。本节还阐述了 Cox 回归。针对寿命表法提出了带有或者没有因子的几个实例。在对 Kaplan-Meier 法的阐述中，介绍了带有/没有因子、带有分层变量并且针对测定置信区间的一些实例。

第 4.7 节首先介绍了 Cox 模型的特点（SPSS 过程命令 COXREG），然后将这种方法与寿命表法、Kaplan-Meier 法和线性回归相比较，计算和解释了 Cox 回归的几种变型（时间独立协变量、时间相依协变量、交互作用和“模式”）。接下来的几段介绍了检验 Cox 回归特定前提条件（主要是对截尾、多重共线性和比例性假设的分析）的方法以及如何建立对比（“偏差”、“简单”、“Helmert”等）。最后归纳了所介绍方法的各种前提条件，以及对其进行

检验的方法。

第 5 章借助于 SPSS 分析示例介绍了回归分析方法的其他用途。

第 5.1 节阐述了两种形式的偏回归。第 5.1.1 节介绍了部分最小平方回归（Partial Least Squares, PLS）。尤其是在有很多预测变量、预测变量相互高度相关，并且（或者）预测变量的数量超过个案的数量时，建议使用 PLS。PLS 兼具主成分分析和多元回归的特点，从而可以将任意测量水平、任意数量的（潜）变量之间的因果关系模拟成线性的结构化方程模型。此外，PLS 还支持混合回归模型和混合分类模型。自变量和因变量既可以是定距的，也可以是定类的。从 SPSS 16 版开始提供了 PLS 命令，PLS 是基于 Python 扩展的。第 5.1.2 节介绍了利用 SPSS 过程命令 REGRESSION 进行相关分析的一种偏回归形式。

第 5.2 节介绍了如何利用线性混合模型（SPSS 过程命令 MIXED）对个体生长曲线进行线性建模。个体生长建模（individual growth modeling）大致上也可以改写为“对个体进行重复测量的方差分析”。对于“普通”线性回归而言，只有一条回归线（例如，回归线也会利用轮廓图生成重复测量的方差分析）通常不适合各个不同的个体（线性）运行曲线。但是在进行重复测量的回归分析或者方差分析之前，利用随机截距模型进行建模，就可以根据截距、斜率和两个参数同时估计出个体的运行曲线。借助于一个分为三级的实例分析，下文演示了某个培训项目所有学员的成绩在经过一段时间培训后是否以及在多大程度上有区别。在这个实例中具体检验了：（a）培训学员的（成绩）水平是否波动（截距），（b）培训学员成绩提高的幅度和速度是否不同（斜率），以及（c）在考虑到培训学员成绩水平的情况下，他们成绩的提高幅度是否不同（两个参数）。

第 5.3 节介绍了岭回归（SPSS 宏“Ridge-Regression.sps”）。岭回归可以（主要是通过目视）检验的是，可能具有多重共线性的数据是否可以用多元线性回归分析来进行分析。与其他统计方法相反，SPSS 岭回归没有采用菜单导航，而是只能采用宏的形式。但是，岭回归的实施并不复杂。本节主要演示了多重共线性的可视化，以及如何针对所选择的 K 值计算岭回归。由于 2008 年的 SPSS 16 版没有宏“Ridge-Regression.sps”，因此本节的实例主要基于 SPSS 15 版的宏。

第 6 章用一个总览表介绍了利用 SPSS 进行回归分析的其他方法（例如，多个因变量的回归分析）。与前面的章节相反，没有对示范性的 SPSS 分析进行复核。这个总览对完全性不做要求。作为例证，这里归纳了示范性分析的、没有注解的语法示例，主要是因为目前只有通过这种方式才能看到相应的要求。

为了评估 SPSS 的输出结果，了解其统计定义和推导过程是必不可少的。在本书的最后一章，归纳了一些最重要的统计方法的公式。

在开始进行分析之前，请先确定你的数据足以进行分析。检查你的数据是否有潜在的错误（主要是完整性、统一性、缺失值、离群值、重复值）。信任是非常好的，但是检查更加重要。对数据质量的准则和用 SPSS 验证数据质量这方面知识感兴趣的读者，请参考 Schendera（2007）的著作。

在这里要特别感谢以下各位的专业建议和（或）他们通过语法、数据和（或）资料对本书做出的贡献：Vijay Chatterjee 教授（西奈山医学院，纽约，美国）、Mark Galliker 教授（伯尔

尼大学，瑞士）、Jürgen Janssen 教授（汉堡大学）、Mitchel Klein 教授（埃默里大学罗琳斯公共卫生学院，亚特兰大，美国）、Roderick J.A. Little 教授（密歇根大学，美国）、Daniel McFadden 教授（加州大学伯克利分校，美国）、Rainer Schlittgen 教授（汉堡大学）、Stephen G. West 教授（亚利桑那州立大学，美国）、Matthew M. Zack（疾病控制中心，佐治亚州亚特兰大，美国）。

还要感谢德国 SPSS 软件慕尼黑有限责任公司的 Alexander Bohnenstengel 先生、Sabine Wolfrum 女士和 Ingrid Abold 女士慷慨地提供了这套软件和相关的技术资料。同样，也要感谢 SPSS 瑞士分公司的 Josef Schmid 先生和 Daniel Schloeth 博士。

感谢奥登伯格出版社 Schechler 博士对发表本书的信任以及对此提供的大力支持。Peter Bonata 先生（科隆）为 Cox 回归一章奠定了基础。Volker Stehle 先生（埃平根）负责本书的印刷排版工作。Stephan Lindow 先生（汉堡）为本书制图。Markus Schreiner 先生（海德堡）为特殊分布提供了随机数据。如果本书中还有阐述不清楚或者错误的地方，欢迎各位读者不吝赐教。

Dr. Christian FG Schendera（克里斯蒂安·FG·申德拉博士）

瑞士，伯尔尼

目 录

第 1 章 相关	1
1.1 引言	1
1.2 第一个前提条件：尺度水平	4
1.3 其他前提条件：线性、同方差性和连续性	5
1.4 说明：对线性的图形检验	6
1.4.1 过程 GRAPH, Scatterplot 选项	6
1.4.2 SPSS 过程命令 CURVEFIT	7
1.5 相关系数的统计和解释	10
1.5.1 相关系数的统计量	11
1.5.2 相关系数的解释	11
1.6 利用 SPSS 的计算（示例）	14
1.7 难点：线性、产生错觉相关和一型误差累积	16
1.7.1 产生错觉相关和偏相关	16
1.7.2 一型误差累积问题	19
1.8 特殊用途	20
1.8.1 相关系数的比较	21
1.8.2 比较相关的一致性	22
1.8.3 正准相关	23
1.9 计算皮尔逊相关系数的前提条件	24
第 2 章 线性回归和非线性回归	26
2.1 线性回归：有因果方向的关联	27
2.1.1 双变量线性回归：利用 REGRESSION 的回归分析概述	27
2.1.2 双变量线性回归的示例和语句——第一步：根据杠杆值和残差检验线性 并识别离群值	32

2.1.3	输出结果和解释	37
2.1.4	过程 2: 删除离群值的效应——选出的输出结果	49
2.1.5	说明: 绘制回归直线 (IGRAPH) 的图形	51
2.2	非线性简单回归	51
2.2.1	利用线性回归对线性函数进行分析	53
2.2.2	利用线性回归分析调查非线性函数	53
2.2.3	将非线性函数线性化, 并利用线性回归进行调查	54
2.2.4	利用非线性回归分析非线性函数: 非线性回归	56
2.2.5	更高的要求: 带有两个预测变量的非线性回归	67
2.2.6	用于非线性回归的 SPSS 过程 NLR 和 CNLR	70
2.2.7	非线性回归的假设	73
2.2.8	总览表: 非线性回归的模型	74
2.3	多元线性回归: 多重共线性和其他难点	76
2.3.1	多元回归的特点	77
2.3.2	第一个例子: 多元回归特殊统计的解释	79
2.3.3	第二个例子: 多重共线性的识别和消除	93
2.4	计算线性回归的前提条件	99
第 3 章	逻辑回归和有序回归	105
3.1	引言: 因变量的因果模型和测量水平	106
3.2	二元逻辑回归	107
3.2.1	逻辑回归方法和与其他方法的比较	107
3.2.2	示例界面和语法: 逐步法 (BSTEP)	111
3.2.3	输出结果和解释	114
3.2.4	示例和语法: 直接法 ENTER	122
3.2.5	输出结果和解释	123
3.2.6	补充说明逻辑回归的理论检验 vs 诊断: 模型拟合优度 vs 预测效率	127
3.2.7	二元逻辑回归的前提条件	127
3.3	有序回归	133
3.3.1	有序回归方法和与其他方法的比较	134
3.3.2	例 1 界面操作和语法: 定距预测变量 (WITH-选项)	135
3.3.3	输出结果和解释	138
3.3.4	例 2 和语法: 分类预测变量 (BY 选项)	143
3.3.5	输出结果和解释	144
3.3.6	有序回归的前提条件	151
3.4	多项逻辑回归	152
3.4.1	例子、界面选择和语法: 主效应模型 (二元因变量)	153
3.4.2	输出结果和解释	159

3.4.3	补充说明：逐步计算带有一个二元因变量的模型： NOMREG REGRESSION 和 LOGISTIC REGRESSION 输出结果的比较.....	163
3.4.4	特殊情况：带有定量预测变量的巢式病例对照研究（1:1）——示例、 语法、输出结果和解释	164
3.4.5	补充说明：LOGISTIC REGRESSION 对比 NOMREG （区别）	168
3.4.6	多项逻辑回归的前提条件	169
3.5	本章所介绍的各种回归方法的比较.....	173
第 4 章	生存分析	175
4.1	生存分析概述	176
4.2	生存分析的基本原理	178
4.2.1	生存函数 $S(t)$	178
4.2.2	确定生存函数 $S(t)$	179
4.2.3	其他函数	180
4.3	截尾数据	182
4.3.1	非期望事件或者未发生目标事件	182
4.3.2	对截尾数据与非截尾数据做不同处理的三个理由.....	183
4.3.3	失效数据和截尾的处理（三种方法）	184
4.4	估计生存时间 $S(t)$ 的方法.....	185
4.4.1	保险精算法和寿命表法	185
4.4.2	使用 Kaplan-Meier 法估计生存时间 $S(t)$	186
4.4.3	无截尾和有截尾的示例（方法：Kaplan-Meier）	187
4.5	对多个组进行比较的检验.....	190
4.6	利用 SPSS 进行生存分析.....	192
4.6.1	示例：无因子 Kaplan-Meier 法	193
4.6.2	示例：采用因子的 Kaplan-Meier 法	198
4.6.3	利用因子变量与分层变量进行比较（Kaplan-Meier 法）	202
4.6.4	Kaplan-Meier 分析的置信区间	207
4.6.5	不带因子的寿命表法计算示例	209
4.6.6	带有因子的寿命表法计算示例	212
4.6.7	计算生存分析的首要条件	216
4.7	Cox 回归.....	218
4.7.1	Cox 模型简介和背景知识.....	218
4.7.2	带有定量协变量的 Cox 回归	222
4.7.3	带有二元协变量的 Cox 回归（ $k=2$ ）	230
4.7.4	带有分类协变量的 Cox 回归（ $k>2$ ）	233
4.7.5	针对交互作用的 Cox 回归.....	237
4.7.6	检验 Cox 回归的前提条件.....	249

4.7.7 带有时间相依的定量协变量的 Cox 回归	256
4.7.8 Cox 回归的特定前提条件	261
4.7.9 附录：对比方法	264
第 5 章 回归分析的其他应用实例	269
5.1 偏回归	270
5.1.1 运用 PLS 过程 (Python Extension) 进行计算	271
5.1.2 运用 SPSS 过程 REGRESSION 进行计算	278
5.2 个体生长曲线	281
5.2.1 方法 1：随机截距模型	282
5.2.2 方法 2：随机斜率模型	286
5.2.3 方法 3：随机截距和随机斜率模型	288
5.3 岭回归 (SPSS 宏)	290
5.3.1 利用岭迹实现多重共线性的可视化	291
5.3.2 岭回归的计算	294
5.3.3 SPSS 宏 “Ridge-Regression”	295
第 6 章 其他方法和模型 (一览)	301
6.1 通过 SPSS 菜单调用其他回归方法	301
6.2 可用语句调用的其他回归形式	308
附录 A 公式	309
参考文献	320
您对本书的建议和意见	327
作者简介	328

第 1 章 相关

第 1 章介绍了往常被人们低估的相关分析（SPSS 过程命令 **CORRELATIONS**）的入门知识。同时，本章开头部分解释了关联（因果性），并列举了几个错误结论的例子，如人们常说的喜欢玩暴力型电脑游戏和个人暴力倾向之间的关联（参见第 1.1 节）。借助于相关分析，阐述了首要的、对于（线性）回归分析也适用的前提条件，例如，尺度水平、同方差性和连续性（第 1.2 节和第 1.3 节）。接下来的几段阐述了线性、产生错觉相关和一型差误累积几个主题，并且还解释了为什么只给定相关系数的数值是远远不够的（第 1.7 节）。同时，还介绍了对相关系数的统计和解释，以及对线性的图形检验（第 1.4 和第 1.5 节）。然后对利用 SPSS 的相关系数做出了计算和解释（第 1.6 节）。此外，还介绍了其他一些特殊的用途，主要是相关系数的比较和正准相关（第 1.8 节）。最后一段归纳了实施相关分析的各种前提条件以及对其进行检验的方法（第 1.9 节）。如果读者对回归分析感兴趣，则强烈建议先阅读关于相关分析的章节。

1.1 引言

没有因果方向的关联
相关不等于因果关系

科学调查的目的通常是分析两个变量之间的相关性。下列问题是相关分析的应用示例：

- 妊娠期与新生儿体重是否相关？
- 食物的重量（例如，单位：克）和营养成分（例如，单位：焦耳）是否有关联？
- 汽车的发动机功率和耗油量是否有关联？

在进行相关分析时，首先应注意一条重要的基本原则：相关不等于因果关系！“相关无法证实因果关系”（Pedhazur, 1982, 579）。相关分析可以说明两个变量之间是否以及在多大程度上存在关联，但不能说明其关联的类型，即无法反映出这两个变量（如果有的话）中哪一个是原因，哪一个是结果。但是这适用于相反的情况，也就是说，如果不存在双变量相关，则没有双变量因果关系。

例如，如果观察到两个变量 *A* 和 *B* 之间具有统计学上的重大关联，则原则上可能有四个因果解释（参见 Pedhazur, 1982, 110ff., 578ff.）：

- *A* 影响 *B* 构成因果关系
- *B* 影响 *A* 构成因果关系
- *A* 和 *B* 受第三个或者多个变量的影响构成因果关系
- *A* 和 *B* 相互影响（构成因果关系）

相关系数无法阐明哪个因果解释是正确的。两个变量之间的相关是因果关系的必要条件，但不是充分条件。

因果模型中的哪一个是最可信的（原则上还可以设想很多其他因果模型），不是由 *A* 和 *B* 之间的相关，而只能是用一种恰当的理论来确定。只有逻辑和可靠的结论是解释相关的坚实基础。

即使在公开出版物中，也经常将相关与因果关系混淆，如果作者自己没有注意到这个问题，则读者也应予以批判地看待。

示例 1：Gale 等人（2006）的著作报道了一个现象，即童年时的高智商与成年时成为素食主义者的高概率之间存在关联。在这里就混淆了相关和单向因果性，因为：（a）如果有人是素食主义者，就是因为他们聪明吗？（b）或者是否有人仅仅因为重视健康饮食而成为素食主义者，而这不一定与高智商有任何关系？

示例 2：一本关于癌症研究的出版物也提出了类似的论证。在使用荷尔蒙疗法期间，服用荷尔蒙会恶化而不是抑制女性的乳腺癌。但是根据 Peter Ravdin（2006）的阐述，乳腺癌确诊病例的数量如今出现了下降，因为在美国有越来越多的妇女放弃了荷尔蒙疗法。大量妇女中断了荷尔蒙治疗，从而产生了这样一个效应，即乳腺癌病例的数量在短短几个月内快速上升。实际上，尽管相关等同于因果关系看起来具有相当的可信度：即使荷尔蒙疗法和乳腺癌风险上升事实上存在关联，但是并不能自动就得出相反的结论，即减少荷尔蒙疗法的应用（例如，通过修改开处方的惯例）肯定会使乳腺癌风险降低。服用荷尔蒙不是唯一的致癌因素。

示例 3：Cha 等人（2001）在著名的《生殖医学杂志》上发表了一篇文章，说看起来经验证明了祷告和怀孕概率之间存在关联。他们声称，如果由一个祷告小组对不育妇女做祷告，则她们的怀孕概率是没有接受祷告的妇女的两倍之多。这项研究被人们在很长一段时间内称为“随机临床评价研究”，主要是因为妇女们根本不知道有人在为她们祷告，并且祈祷者与她们相隔了几千公里。Cha 等人（2001）关于祷告和怀孕概率之间关联的文章是纯粹的谎言，其中一个作者因为多次诈骗已经受到法律制裁，甚至无法证明曾经做过这项调查研究。简单的重复并不能证实祷告和怀孕之间具有所声称的关联。但是，这并没有阻止盲目听信的公众将这项研究作为所谓的“信仰疗法”的疗效而予以大肆宣扬。

示例 4：在媒体反馈的研究领域，人们常说喜欢玩暴力型电脑游戏和游戏玩家暴力倾向之间存在关联，其实人们是从几个层面上混淆了原因和关联。实际上，这个“命题”本身就是科学上没有正确地，特别是没有以相关系数形式解释关联的很好例子。

层面 1：将双侧相关减少到一个单侧因果方向。

将所声称的（只是隐式双侧的）、喜欢玩暴力型电脑游戏和个人暴力倾向之间的关联减少到喜欢玩暴力型电脑游戏作为暴力倾向的一侧原因。在这里就错误地把第二个潜在的作用方向排除在外，即个人暴力倾向可能是沉迷于暴力型电脑游戏的原因；这绝不能支持第一个作用方向（即喜欢玩暴力型电脑游戏是个人暴力倾向的唯一原因）的论点的可信性。

层面 2：将复杂的关系网减少到唯一一个原因（单向因果性）。

在所声称的唯一一个作用方向（即暴力型电脑游戏作用于暴力倾向）中，将喜欢玩暴力型电脑游戏具有单向因果性这个论点隐含地表述为唯一一个影响因素。在社会科学领域，这样的观点就会被人们视为落伍（过时）的或者太过简单化。一个关联（如相关）仅仅能够将两个变量代入一个模型，这并不能说明这个模型是否正当地反映了（实证）现实的复杂性。

层面 3：由于不现实的简单论点远离（实证）现实的复杂性。

只有在个别情况下，才能用单向因果和单调的变量影响描述社会科学领域研究对象（实证）现实的复杂性。换言之，在这个领域通常认为各种因素形成了一个复杂、动态的网络。鉴于已卖出或者人们正在玩的电脑游戏的情况，以及例如杀人狂的数量，从电脑游戏消费的实际情况来看更应得出下列观点：几乎每个杀人狂都玩过暴力型电脑游戏，但不是每个暴力型电脑游戏玩家都会大开杀戒。再仔细观察一下，就可以看出，前面声称的、喜欢暴力型电脑游戏和玩家暴力倾向之间的关联至少有两个错误。

一个是取样错误。将比例很小的极端组（如杀人狂）当作事实结果，以其作为部分（“pars”）推断出了所有玩家总体（“toto”）。但是由于取样错误，这个极端组并不能代表整体。此外，这里还错误地使用了“以部分代整体”原则。

另一个是思维错误。关于杀人狂（他们也玩暴力型电脑游戏，但是从数量上来看明显是少数）变得具有暴力倾向的原因的命题，在很大程度上忽视了一个问题，即为什么绝大多数暴力型电脑游戏的玩家没有大开杀戒。结论是十分明确的：（a）因为作用方向可能完全相反，从社会心理学角度来看，游戏玩家的心理状态不仅仅决定了他们是否喜欢玩暴力型电脑游戏，而且还决定了是否会（自主地）实施暴力行为；（b）因为如此一来，暴力型电脑游戏的单向因果效应这个论点就无法立足了。对电脑游戏下禁令改变不了前面所述的伪科学命题，因为相关不等于因果关系。

在如今的媒体心理学研究中，找不到证据说明电脑游戏（如魔兽世界）和暴力行为之间存在直接联系。与此相反，媒体的偏见、意外的效果（如在玩暴力型电脑游戏后感到放松）以及一些起到中和作用的因素（例如，年龄、性别、个性和社会结构）却经常见诸报端。也可以利用复杂的统计方法，例如，多元回归“构建”因果关系模型。这样的因果关系结构通常是很难让人一目了然的，也就导致一些伪科学的“发明”，得出夸张的结论，但是对所依据的数据进行仔细分析就会发现，这些结论是完全站不住脚的。这种“发明”的典型例子就是，例如，有人声称，死刑或者允许佩戴武器降低了犯罪率（参见 Goertzel, 2002）。

两个变量之间的相关不等同于两种构件（construct）之间的关联，只是表明了一些观点。例如，操作性定义或者样本依存性。相关分析也可以理解为，对无须太多解释的简单理论的具体操作方法进行统计建模。相关分析中具有统计显著性的事件并不排斥其他竞争性模型的有效性。因此，将相关分析解释为单向因果时，就同时犯了以下几个错误：

- 将一个变量等同于一个构件
- 将相关减少到一个因果方向
- 将复杂的关系网减少到唯一一个原因（单向因果性）

因此，要谨慎地看待相关分析的“证据”。例如，所声称的两个变量之间的关联确实表现出统计显著性（也与显著性的绝对化相关，参见，Witte, 1980; Schendera, 2007）。

如果要检验在一个因果模型中，是否一个变量会有规律地造成另一个变量的变异，则可以选择回归法代替相关法。对于复杂的模型，可以考虑使用的方法主要是偏相关（第 5.3 节）或者偏回归，必要时也可以使用路径分析（第 5.1 节）。

为了测定两个变量之间关联的程度，统计学发展出了很多测量方法。然而，在各种文献中，对于相关量度（“相关”、“相联”）的专业术语并没有取得统一（参见 Lorenz, 1992, 58ff.）。例如，如果调查几列数值对之间（线性）关联的强度，则人们就将定距型数据或者定序数据（量度数据、极差数据）称为具有相关性。对于交叉列表、 2×2 表或者列联表则使用相联，或者列联的叫法。对于定序变量，则根据斯皮尔曼相关分析方法将关联强度称为相关。相反，根据肯德尔或萨默斯相关分析方法则称为列联或者相联。Bortz（1993，参见第 6.3 节）甚至对二元变量采用了相关的说法。选择使用哪种方法，最终取决于变量类别的数量、分布和尺度水平（参见关于列表分析的章节）。在某些情况下，是否存在一个因果模型（例如，“X 造成 Y”）、关系的原因、变量的数量和其他因素也很重要。但是，所有方法都遵循一个基本原则。从根据经验观察到的关联和理论上的最大关联两个方面对变量进行比较。换言之，将各个变量当前实际的共同点，与变量之间关系达到完美时变量之间本应具有的共同点进行比较。

皮尔逊相关系数（又称协方差相关或 Bravais-Pearson 相关）描述了两个定比、线性相关的变量（测量值序列）之间不受其单位影响的关联强度（又称紧密度）。

1.2 第一个前提条件：尺度水平

数据的尺度类型决定了用何种形式证明两个变量之间的（线性）关联。如果数据是定距型，则可以将皮尔逊相关系数作为量度，利用肯德尔或者斯皮尔曼相关分析方法分析连续定序变量；如果两个定距变量之间的关联不是线性，而是单调的，并且这条信息是充分的，则也可以使用这些方法（如斯皮尔曼极差相关）。

如果有离散尺度的定序变量，则可以使用如伽玛、萨默斯等量度（具体可参考列表分析，例如，Schendera, 2004，第 12 章）。

如果是两个分类（“定性的”）变量，则可以用列联系数描述其间的关联。所选择的相关分析方法要与所调查变量的尺度水平一致，这一点也很重要。对于尺度水平不同的成对变量，

应始终选择使用较低尺度的变量的尺度水平。

关联性和相关性度量一览表

	区间	有序	名义
区间	皮尔逊相关系数		Eta ² (R ²)
有序		皮尔逊相关系数（如果距离相同），斯皮尔曼相关系数 伽玛，肯德尔 τ -b 系数，肯德尔-斯图尔特 τ -c 系数，萨默斯 d 系数	
名义			ϕ 系数 Cramer V 值 列联系数 λ 不定性系数 列联系数 Cochran Q

这个表格中都是对称量度，只有 Eta²例外。在计算相关性或者关联性时，哪个是自变量、哪个是因变量并不重要。例如，当自变量是定距的、因变量是定类的时，在这种情况下就可以使用 Eta²。重要的是，接下来所选择的统计量与所确定的测量值分布形状（函数）一致。对于两个比例量度的相关性，Cohen 等人（2003，60-63）建议要小心谨慎，甚至完全不将其相关。

接下来介绍皮尔逊相关分析。这种分析方法的前提条件是，两个变量至少是定距的，但是如前所述，两个所调查变量之间的关联是线性的。

1.3 其他前提条件：线性、同方差性和连续性

还需仔细探讨的一个皮尔逊相关系数前提条件是，两个所调查的变量（或者更准确地表述：其成对测量值）的图形表现出线性关联。例如，在散点图中，成对测量值的排列“基本上”呈线性。如果一个散点图中的数据呈线性排列，则可以选择线性相关的量度。

如果两个变量精确地具有同样的分布形状（但是不一定是正态分布），那么就达到了最大程度的正关联（ $r=1.0$ ）。如果两个变量精确地呈现镜像相反的分布，则达到了最大程度的负关联（ $r=-1.0$ ）。为了利用皮尔逊相关系数描述双变量关联，变量不一定必须是正态分布的。两个变量的分布呈现的相互距离越大，关联就越小（Cohen 等人，2003，53）。

仔细观察就可以发现，线性结合了三个可以通过图形（如散点图）检验的特征。

- **成对测量值的线性排列：**通过皮尔逊相关系数，无法恰当地描述曲线的测量值分布。
- **围绕直线的发散程度：**发散程度越窄，相关系数越大。排列越发散、越像是云状，相关系数越小。
- **删除离群值：**不存在离群值。无论是朝向函数方向，还是垂直于函数的离群值都会使相关系数产生偏误（参见 Schendera，2007）。

出现较高相关系数的前提条件是数值分布呈线性，因此相关系数也可以不是线性检验的一种方法。如果没有呈现出三个特征中的至少一个，例如，线性函数、发散程度最小、离群值最少，就会产生较小的相关系数。由于只要有一个离群值（如果没有就会呈现完美的线性相关）就会在很大程度上使估计过程产生偏误，从而产生较低的相关系数，因此在这里也可以不将相关系数称为线性检验的方法。

如果散点图显示数据走向呈曲线形，则不能选择线性关联的量度，而要选择非线性关联的量度。换言之，关联程度统计量的标准量度和皮尔逊相关系数，都以数据呈线性排列为前提条件。利用双变量散点图可以方便地检验是否存在“线性”，因此是一种“对线性的图形检验”。

1.4 说明：对线性的图形检验

1.4.1 过程 GRAPH, Scatterplot 选项

散点图描绘了在一个坐标系中的成对测量值（“散点图”）。散点图适合用来以图形方式描绘至少两个定距型定量变量之间的关系。在散点图中还经常插入（回归）直线，表示两个变量的（曲线形）线性关联（例如，参见下文的 CURVEFIT）。

说明

下面的散点图展现了在两个刻度轴上的两个变量。一个变量确定水平轴，另一个变量确定垂直轴。 X 轴（在这里是腰围）变量的每个数值都与 Y 轴（在这里是体重）变量的相应数值录入到了坐标系中。利用由此产生的散点图，可以表现出两个变量之间的关联。在本例中，两个变量之间呈现线性的正关联（图形是“线性检验”）。由于成对测量值的两种测量值分别发散在假想的相关直线两侧，并且与相关直线的距离基本相等，因此可以同时确定有同方差性（发散程度的一致性）。由于直线没有间断（例如，没有断续分布），如同在分析极端组时可能出现的那样，则确定了直线的连续性。

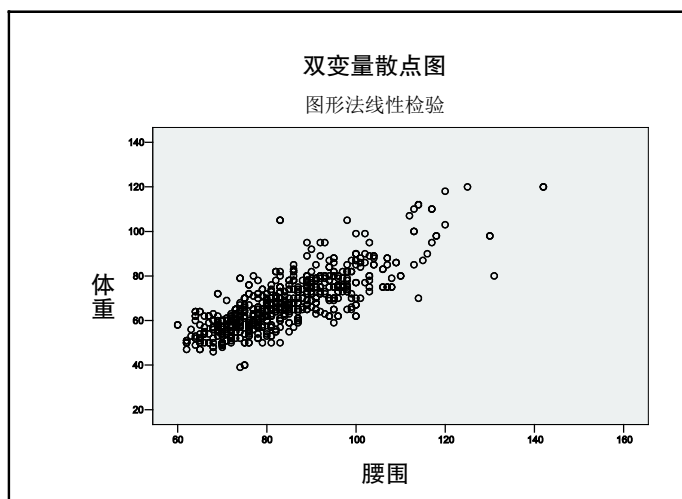
预设定

在 SPSS 程序主界面选择以下菜单项：编辑 → 选项 → “查看器”选项卡。

请确定选项“在日志中显示命令”已选中。

在 SPSS 程序主界面选择以下菜单项：图形 → 旧对话框（旧版本 SPSS 上没有） → 散点/点状 → 简单分布 → 定义。

单击 X 轴中的 TAILLE（腰围）和 Y 轴中的 GEWICHT（体重）。在“选项”路径下，单击“整行删除”命令来处理缺失值。在“标题”路径下，确定标题和子标题。单击“确定”按钮，调用图形。



```
GRAPH
/SCATTERPLOT (BIVAR) =
    taille WITH gewicht
/MISSING=LISTWISE
/TITLE= "双变量散点图"
/SUBTITLE= "图形法线性检验".
```

解释说明：

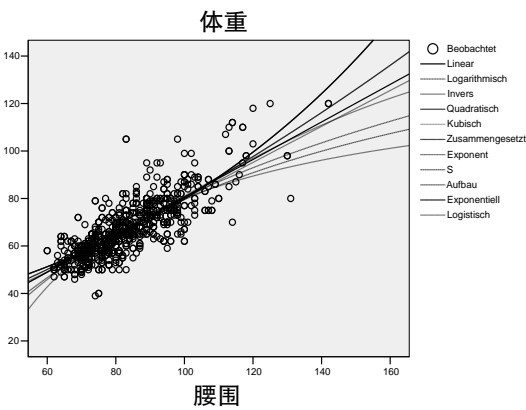
GRAPH 命令调用一个图形。利用 /SCATTERPLOT (BIVAR) 命令确定类型，即双变量散点图。在等号后面是两个变量的名称，其数值对需要录入到散点图中。提到的第一个变量（在这里是 TAILLE，腰围）在 X 轴上表示。GEWICHT（体重）在 Y 轴上表示。根据 MISSING= 可以确定应如何处理可能出现的缺失值。这里选择了选项 LISTWISE 后就执行整行删除（还可以通过 VARIABLEWISE 命令逐个变量地删除个案）。通过 TITLE、SUBTITLE 和 FOOTNOTE（必要时）命令输入标题、子标题和脚注的文字。

1.4.2 SPSS 过程命令 CURVEFIT

线性关联是达到较高相关系数或者计算皮尔逊回归的基本条件之一。SPSS 过程命令 CURVEFIT 提供了另一种检验两个变量之间是否存在线性关联，或者另一个函数是否能更好地解释这种关联的方法。

说明

SPSS 过程命令 CURVEFIT 的功能远远超过双变量散点图的简单排列。CURVEFIT 不仅检验可能存在的线性关联，而且还检验另外 10 个关联模型（主要是指数、指数分布、逆、立方、对数、二次、S (S)、增长和复合）。从根据经验存在的成对测量值（“观察”）的排列中，CURVEFIT 命令截取测算出的直线函数（如果一次性绘制大量函数，就有可能看不到全貌）。此外，CURVEFIT 命令针对每个函数都测算出统计参数，例如 R^2 等。因此，不仅可以通过肉眼观察直线，还可以根据统计参数对不同的曲线模型进行比较。此外，通过比较还可以相对简单地判定哪个函数可能比直线模型更好地反映出所调查的两个变量之间的关联。



Beobachtet 观察 Linear 线性 Logarithmisch 对数 invers 逆 Quadratisch 二次 Kubisch 立方
zusammengesetzt 复合 Exponent 指数 Aufbau 结构 exponentiell 指数分布 Losistisch 逻辑

在 SPSS 程序主界面选择以下菜单项：分析 → 回归 → 曲线估计 → 定义。

将 GEWICHT（体重）确定为因变量，TAILLE（腰围）确定为自变量。从下面提供的模型中选择想要的曲线函数。勾选“在等式中包含常量”和“根据模型绘图”。单击“确定”按钮，调用曲线估计。



```
TSET NEWVAR=NONE .
CURVEFIT
/VARIABLES=gewicht WITH taille
/CONSTANT
/MODEL=
  LINEAR LOGARITHMIC INVERSE
  QUADRATIC CUBIC COMPOUND
  POWER S GROWTH EXPONENTIAL
  LGSTIC
/UPPERBOUND=150
/PLOT FIT .
```

语句说明

CURVEFIT 命令调用了曲线估计的方法。CURVEFIT 命令标准化地输出一个曲线估计图和回归统计量的综合表，其中主要包括曲线函数或曲线方法、 R^2 、自由度、F 值、显著性水平、上限（Upper bound）、常量（b0）和回归系数（b1, b2, b3）。置信区间预设定为 95%。CURVEFIT 标准化地整行删除缺失值。

在 VARIABLES 后面是曲线函数的两个变量的名称，数据都应针对这些曲线函数进行调整。提到的第一个变量（在这里是 *taille*，腰围）作为自变量被列入建模，只能给定一个自变量。然后提到的第二个变量（在这里是 *gewicht*，体重）构成了因变量，可以给定几个因变量。如果 *taille*（腰围）成为因变量，*gewicht*（体重）成为自变量，则结果会得出其他函数（对此参见简单线性回归的说明），但 R^2 还是同一个。只能给定一个 VARIABLES 命令。/CONSTANT 命令决定了回归方程是否应包含一个常量（或者是否不包含一个常量：NOCONSTANT）。根据 /MODEL= 命令，一次性可以给出最多 11 个不同的回归模型（也可以通过选项 ALL）实现。由于 CURVEFIT 命令针对每个因变量和模型曲线都自动创建四个新的变量，因此在数据集很大时不应再将其作为所需的模型曲线调用。下面详细介绍不同的回归模型。如果调用一个逻辑回归模型（LGSTIC），则必须利用 /UPPERBOUND 单独给出一个上限值，这个值是正数并且大于所给出的所有因变量中的最大值；对于现有数据，已经给出了数值 150。针对逻辑回归模型，应在输出结果中给出这个上限值。利用 PLOT=FIT（预设定）命令调用曲线估计图，PLOT=NONE 删除曲线估计图的输出结果。

在 CURVEFIT 前面的 TSET NEWVAR= 确定了用于处理因变量数值的预设定。例如，时间序列和序列变量。如果是 NONE 命令，则不存储新的变量。相反，在 CURRENT 命令（预设定的）和 ALL 命令时存储变量，即在 CURRENT 命令时替换现有变量，相反在 ALL 命令时不替换。

预设定的例子不适用于时间序列数据。下述输出结果经过简化。

模型总结和参数估计值

因变量：体重（Gewicht）

方程	模型总结					参数估计值			
	R^2	F	自由度 1	自由度 2	显著性	常量	b1	b2	b3
线性（Lin）	0.645	1247.205	1	686	0.000	1.532	0.791		
对数（Log）	0.632	1176.897	1	686	0.000	-232.420	67.965		
逆（Invers）	0.607	1058.998	1	686	0.000	136.157	-5604.780		
二次	0.646	625.993	2	685	0.000	14.652	0.490	0.002	
立方	0.646	625.993	2	685	0.000	14.652	0.490	0.002	0.000
复合	0.633	1183.187	1	686	0.000	26.405	1.011		
幂函数	0.633	1184.755	1	686	0.000	0.958	0.961		
S 型（S）	0.620	1120.686	1	686	0.000	5.176	-79.995		
结构函数	0.633	1183.187	1	686	0.000	3.274	0.011		

续表

方程	模型总结					参数估计值			
	R ²	F	自由度 1	自由度 2	显著性	常量	b1	b2	b3
指数	0.633	1183.187	1	686	0.000	26.405	0.011		
逻辑	0.643	1236.631	1	686	0.000	0.051	0.978		

自变量是腰围（taille）

输出结果说明

如上所述，绘出曲线估计图后，在根据经验存在的成对测量值（“观察到的”）的排列中，针对每个调用的函数在散点图中绘出一条直线（线性的、对数的等）。如果将各条直线的比较限制在基本数据的值域内（SPSS 使绘出的直线超过作为基础的值域，这实际上是不允许的），则可以根据曲线估计图的图形确认，哪几个测算出的函数将会得出几乎同样的结果。注意，不要将这里的“同样”结果与“同样好”的结果混淆。为了更有说服力地在函数之间做出选择，可以查询表格中的回归统计量。

表“模型总结和参数估计值”列出了所调查模型的方程和特征值，因变量“体重”和自变量“腰围”各自在表格的上方和下方给出。在“方程”、“模型总结”和“参数估计值”下面列出了曲线函数的方程和相应的“R 方”[R²]、F 值 [“F”]、自由度 [“1”或者“2”]、显著性水平 [“Sig”]和常量（b0），从“b1”开始根据不同模型列出了不同的回归系数。与先前的 SPSS 版本相反，在另一个表格中输出了所给出的允差上限，以及所遵守的或者没有遵守的允差。可以通过 TSET 设置 QUA 和 CUB 的允差标准。

根据参数选择曲线函数的前提条件是，利用曲线估计图可以认为所绘出的直线展现了对观察到的成对测量值的合理估计。如果测定的曲线完全没有反映出经验分布，那么再好的参数也是毫无意义的。下一个标准是显著性。首先只观察 F 检验达到显著性的曲线函数（在本例中是所有模型）。下一个标准是 R²。从显著的模型中，只观察 R²值最高的曲线函数（在本例中是 LIN、QUA 和 CUB）。下一个标准是回归方程是否简单。因此，更大程度的方差解释就需要更复杂的回归方程，但是无法总是用方差解释来更好地说明复杂的回归方程的存在合理性。例如，如果线性和三次回归模型的 R²之间只有 0.001 的区别，但是为此付出的代价就是，在二次函数的方程中有两个额外的变量（参见总览表），则应优先采用更简单的方程，在这里就是线性回归函数。这样做的优点是，不仅可以继续利用线性相关模型或者线性回归模型进行计算而没有实质性的信息丢失，而且利用线性模型对计算结果或者模型的解释，比利用二次模型的解释更加简单。但是最后应再次确认，用图形和统计方法找到的曲线函数是否真正适合描述相互关联的构件。为了进行最后的观察，通常应进行两方面的试验：（a）如果延长所找到的函数超过现有数据的值域，则可能会导致什么样的后果；（b）根据值域不同，公共函数是否可以分解为单个的、可能相互相通的函数。

1.5 相关系数的统计和解释

如何计算和解释皮尔逊相关系数？

1.5.1 相关系数的统计量

为了计算皮尔逊相关系数，可以使用几个不同类型的公式和计算方法。

$$r = \frac{S_{xy}}{\sqrt{S_x^2 \cdot S_y^2}} \text{ 或 } r = \frac{\text{COV}(x, y)}{S_x \cdot S_y} \text{ 或 } r = \frac{\sum_i ((x_i - \bar{x}) \cdot (y_i - \bar{y}))}{n \cdot S_x \cdot S_y}$$

$$\text{或 } r = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{((n \sum x_i^2) - (\sum x_i)^2)((n \sum y_i^2) - (\sum y_i)^2)}}$$

例如，在第一个公式中， s_{x^2} 和 s_{y^2} 是变量 x 和 y 的方差， s_{xy} 是 x 和 y 的协方差。根据这个公式，皮尔逊相关系数 r 定义为 x 和 y 的协方差除以 x 和 y 方差乘积的平方根。当两个变量发生相同变化时，系数达到最大（参见下文）。方差指的是采用最小定距的发散程度。另外两个公式是第一个公式的变型，第三个写法表明了测定简单线性回归的相似性。将两个采用最小定距变量的公共发散程度定义为协方差。对于标准化变量，协方差等于相关度。

对于计算作为描述性量度的相关系数 r 而言，不需要假设两个变量的分布状况。但是，计算假设检验或者显著性检验的 r 的前提条件是双变量正态分布（例如，Pedhazur, 1982², 40；参见 Schendera, 2007，如何处理截尾分布）。

解释相关系数的前提条件是两个变量之间存在线性关联（如通过散点图看出）。如果数据点形成一条完美的直线，则相关系数达到最大值 1.0。由于 r 是基于观察值和估计值之间距离的最小平方和，因此 r 在最大程度上与数据相匹配。数据点在（假想）直线周围的（有规律）发散程度越大，误差方差就越大，从而 r 越小。双变量发散程度越呈曲线形，线性相关系数就与指数相比越不适合现有数据的发散程度。取而代之，应选择适当的曲线函数（如三次函数、二次函数等）。

相关系数是什么含义？相关系数 r 是基于标准化（Z 转换）的数值，是一个纯粹的数字，不受两个相互相关变量的测量单位影响。因此， r 的绝对值表示了以 Z 值为单位的两个变量之间的线性关联。 r 值越高，越能更好地通过两个变量的其中一个预测另一个。极差相关、 ϕ 系数以及点二列相关系数 r 只是根据 r 公式的等效计算方法（参见 Cohen 等人著作，2003³，第 2 章）。

相关系数的平方（ r^2 ）表示 x 和 y 的共性方差的分量，或者两个变量之间线性相联的方差分量（“重叠”）。 r^2 也称为决定系数。

$1-r^2$ 表示非共性方差或者两个变量之间非线性关联性的分量（“不重叠”）。 $1-r^2$ 也可以解释为一个变量对另一个变量的预测误差（方差估计误差）。

1.5.2 相关系数的解释

通常，通过相关分析调查两个变量之间是否存在关联。皮尔逊相关系数（又称为简单相关系数，或者简称相关系数）是衡量两个连续变量之间线性关联的量度。如本章开头所述，除了皮尔逊相关系数之外，还有其他的相关量度。可以假设相关系数的值在 $-1 \sim 1$ 之间。系数的正负号表明了关联的方向，其绝对值表明了关联的强度。正相关意味着只要其中一个变量的值升

高，则另一个变量的值也会升高（图形：从左下到右上的上升直线，正号）。负相关则意味着如果一个变量的值增大，则另一个变量的值减小（图形：从左上到右下的下降直线，负号）。相关系数很低，则表明所调查的两个变量之间不存在线性关联。如果相关系数大约为 ± 1 ，则表明两个变量之间存在完美的线性关联。如果接近于零，则两个变量之间没有线性关联。

在具有定距的并且可靠测量的变量这个前提条件下，低相关是由测量值分布的，而不是由测量水平本身造成的。两个变量的相关不等同于两个构件的关联；也有可能至少其中一个构件没有可靠地测量（参见 Cohen 等人著作，2003，53-55）。

针对相关系数的显著性检验同时也受基础样本量的影响。显著性在这里并不总是意味着关联性。比假设检验的显著性更重要的是系数的大小。

此时，对系数的评估在很大程度上取决于所提出的问题。对于在 $0\sim 1$ 之间的相关系数值经常做下列解释。

相关系数 r 的解释说明	
r	解释
< 0.2	相关很小
< 0.5	相关小
< 0.7	相关中等
< 0.9	相关大
> 0.9	相关很大

原则上，这样的“解释辅助说明”忽视了相关系数（精确性）和散点图（差异）可以相互补充的信息。

只有当相关系数较大时，例如，如果函数的极差从实用性角度来看非常大，才可以将线性函数解释为相关大。这既可以从图形上体现为线性排列，也可以用数值表现为相关系数大。例如，不是每个定量很大的相关系数都可以解释为定性很高的相关性。

- 示例关联强度取决于变量各自的分布形状，但是不取决于其在 X 轴和 Y 轴上的位置。例如，如果双变量的相关系数为 0.9 ，则将一个（或者两个）变量的所有数值乘以或者除以一个常量（如 10 ）也不会改变相关系数的数值。因此，数值相等与相关系数可能实质上表达的是完全不同的关联。
- 这同样适用于对各个变量测量值域（发散程度）极差的限制。例如，如果将数据（如 $1\sim 100$ ）的极差限制在一小段（如 $1\sim 10$ ），尽管还会产生线性排列或者线性函数，但是再也不能普遍化地将极差描述为“很大”，也就代表了理论上可能达到的极差。即便这样能得出很大的系数也是如此。

因此，根据提出的问题或者数据位置不同，同样大的相关系数可能具有不同的显著性，就像不同大小的相关系数可能就内容而言完全表示同样的含义一样。

此外，当相关系数处于较低或者中等水平时，就产生了一个问题（尤其是在数据量很大时），即无法通过图形解释散点图模型（例如，根据上下重叠的成对测量值）。形状类似的散点图也经常用不同的相关系数来描述。

但是，如果利用散点图的差异无法精确地解释精确量度（例如，相关系数），则用形容词来描述相关系数就意义不大了。

因此，在这里明确不建议使用表示评价意义的术语，如“大”或者“小”。取而代之，相关系数的值应利用相应变量经验上的、以及理论上可能的极差予以说明。

重要的是应该知道，相关系数的大小不仅仅受调查条件（试验设备）影响，而且还受样本特征的影响（随机特性、特征易变性/代表性和大小）。

相关系数的大小也取决于相应的变异性，也就是两个变量的极差。如果忽视了两个变量中至少一个的极差，则就人为地降低了相关系数。根据用不同样本测得的关联，得出了同一个总体的不同相关系数。代表性抽样的随机特性决定了特定的样本变异，从而也决定了样本特定的相关；如果是没有代表性的抽样，则这个效应表现得更为明显。

在非代表性抽样中，极差经常因节选取样的方式而受到损坏。例如，如果不是根据具有代表性的样本，而是只以大学生为样本调查智力因素，则就应认为，分析中只采用了智力因素总极差中的上面一段，而不是具有代表性的一段。

只有利用足够大的样本才能精确地估计一个总体的相关系数；只有根据足够大的样本，才能比较不同的相关系数（假设这些相关系数来自同一个总体）。例如，Diehl 和 Kohr（1999）建立了一个表格，阐明了相关系数变异与样本量的依存性程度。

相关系数变异的依存性

样本量	相关系数的 90%界限
5	-0.82 ~ +0.82
10	-0.55 ~ +0.55
20	-0.38 ~ +0.38
30	-0.31 ~ +0.31
50	-0.24 ~ +0.24
100	-0.17 ~ +0.17
300	-0.10 ~ +0.10
1000	-0.05 ~ +0.05

这个表格的依据是假设相关为 0.00，表明随着样本量的增大，相关系数的变异逐渐减小。当 $n=10$ 时，90% 的相关系数在 ± 0.55 之间。当 $n=100$ 时，90% 的相关系数在 ± 0.17 之间。当 $n=1000$ 时，90% 的相关系数在 ± 0.05 之间。

或许从开始就应该指出，根据大约为零的相关系数不能得出所调查特征不相关的结论。成对测量值与相关直线的偏差越大，相关系数对低估了双变量线性关系强度的低估程度就越大。所调查的特征可能完全是非线性相互关联的，例如，呈 U 形曲线。利用散点图建立出测量值具体分布的精确图像后，才应计算或者解释相关系数。

在进行相关分析时，应始终注意一条重要的基本原则：相关不等于因果关系！相关分析只能说明两个变量之间是否相关以及在多大程度上相关，但是不能说明其关联的类型，也就是无法反映出这两个变量中哪一个是原因，哪一个是结果。如果要检验在一个因果模型中，是否一

个变量会有规律地造成另一个变量的变异，则可以选择回归法代替相关法。但是这适用于相反的情况，也就是说，如果不存在双变量相关，则就没有双变量因果关系。

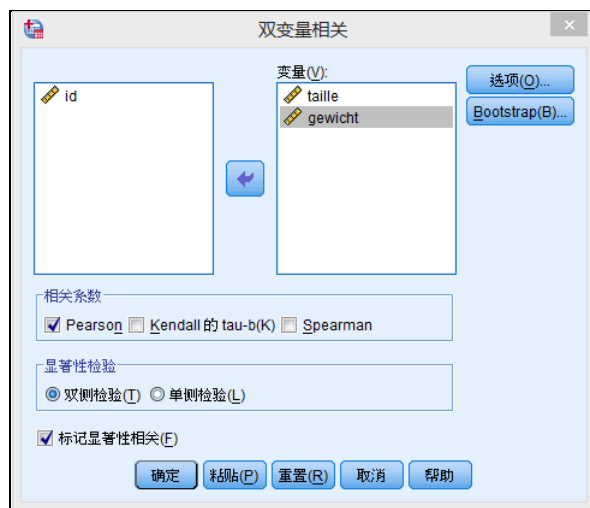
简单（双变量）相关的多变量形式有偏相关、多重相关和正准相关等。例如，偏相关是通过计算删除了由其他的相关变量造成的效应之后，两个变量之间所剩余的相关。

1.6 利用 SPSS 的计算（示例）

针对腰围和体重之间是否具有方向性的关联，下面将计算（双侧）皮尔逊相关的程度。

界面操作

在 SPSS 程序主界面选择以下菜单项：分析 → 相关 → 双变量。



选定和单击选择窗口中的两个变量 **taille**（腰围）和 **gewicht**（体重）。选择“Pearson”选项，选定“Pearson”作为相关系数，在“显著性检验”区域中“双侧检验”单选项。勾选“标记显著性相关”选项。在“选项”路径下，选择“成对删除”来处理缺失值。单击“确定”按钮开始计算。

语句：

```
CORRELATIONS
  /VARIABLES=taille gewicht
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE .
```

CORRELATIONS 命令调用皮尔逊协方差相关的计算。在 **/VARIABLES=** 后面给定对其关联要进行计算的变量名称（在这里是 **taille gewicht**），也可以给定两个以上变量。**/PRINT=** 命令确定显著相关系数的检验方向和标记。通过 **ONETAIL/TWOTAIL** 命令可以给出单侧或者双侧检验方向。尤其是事先不知道关联的方向时，**TWOTAIL**（预设定）适用于双侧检验方向。关联既可以是正关联，也可以是负关联。如果事先已知关联方向或者预计只有某一个方向，则

可以使用 **ONETAIL**。此时，关联只允许是正关联或者负关联。通过 **NOSIG** 命令，可以用星号将显著相关突出显示。 α 为 0.05 的显著值用一个星号标记， α 为 0.01 的显著值用两个星号标记（参见 Schendera，2007，第 19 章，对星号的解释）。用 **SIG** 表示无标记反而更容易实现。通过 **MISSING** 命令确定如何处理缺失值。**PAIRWISE** 命令是成对删除缺失值，**LISTWISE** 命令是整行删除。当只有两个变量时，这两个命令会得出同样的结果；如果超过两个变量，则各自的分析可能会得出差异很大的 N ，尤其是在整行删除缺失值时。此外，通过 **INCLUDE** 命令还可以将用户定义的缺失值引入分析，或者通过 **EXCLUDE** 命令从分析中删除用户和系统定义的缺失值。

除了鼠标控制外，语句还能实现下列功能。

利用 **VARIABLES** 子命令中的关键词 **WITH**，可以计算第一列和第二列所有变量之间的相关。

```
/VARIABLES=VAR1 VAR2 VAR3 with VAR4 VAR5 VAR6
```

此时只测算给定的成列相关，例如，带有 **VAR4** 的 **VAR1**，但是不测算 **VAR1** 和 **VAR2** 之间的相关。有时 **WITH** 选项很适合把某些无法一目了然的相关表缩减为用户感兴趣的成对变量。

例如，用 **/MATRIX** 命令可以将 Pearson 相关系数矩阵写入一个数据集。

```
CORRELATIONS
/VARIABLES= VAR1 VAR2 VAR3 VAR4 VAR5 VAR6
/MATRIX OUT (*) .
```

在本例中，针对变量 **VAR1** 至 **VAR6**，将平均值、标准差和 N 个相关系数写入一个临时数据集。可以用几种方法对这些矩阵继续进行处理，主要包括回归（SPSS 过程命令 **REGRESSION**）、因子分析和聚类分析（**FACTOR**，**CLUSTER**）。

输出结果

相关性			
		腰围	体重
腰围	皮尔逊相关	1	0.803**
	显著性（2 位数）		0.000
	N	711	688
体重	皮尔逊相关	0.803**	1
	显著性（2 位数）	0.000	
	N	688	719

**相关性在 0.01（2 位数）的水平上是显著的。

表“相关性”含有皮尔逊相关系数、显著性以及无缺失值的个案数量（例如，在成对相关时， $N=688$ ）。在一般的相关系数矩阵中，对角线上始终是数值 1，因为每个变量自身都呈现出完美的线性相关。因此，在一般的相关系数矩阵中，相关系数也映射在这个对角线上。

但是，如果利用 **WITH** 选项调用相关分析，则既没有映射出相关系数，表格也不含有第一对角线。在本例中，腰围和体重之间的相关达到 $p=0.000$ 的显著性值和 0.803 的相关系数值。由于没有给出负号，因此这两个变量相互呈显著的、高数值的正相关。

1.7 难点：线性、产生错觉相关和一型差误累积

即使皮尔逊相关从计算角度来看并不复杂，但在对其进行解释时也会经常犯错误。最常犯的错误包括由于寻找显著性而导致的线性、产生错觉相关和一型差误累积。

一个经常犯的错误就是从显著性推导线性。根据显著性无法推导出作为显著性结果的线性关联。相反，线性是允许进行相关分析的前提条件。在这里再强调一遍：如果两个变量相互之间关系紧密，但是其关联是非线性的，则皮尔逊相关系数就不适合作为测量其关联的统计量。简单地说，线性是前提条件，不是显著性的结果。

同样应注意，尽管计算得出的相关从统计学角度来看很高，但是相反作为纯粹的统计关联而无法直接向回推演得出经验关联的现实情况。得出双变量显著相关的结论，通常只是因为没有考虑到第三变量的影响。

例如，如果发现读者年龄和某种日报的价格之间存在高度相关，则尽管两个变量之间明显不存在直接关联，但仍然得出这个错误的结论，这完全是因为没有考虑到第三变量，即造成这个虚假关联的物价上涨率。

对于双变量相关而言，将虚假关联与经验关联混淆的风险是存在的。如何检验干扰变量的影响方法，在下文将通过偏相关予以阐述。

另一个风险是人们经常说的“寻找显著性”。从毫无计划创建的相关系数矩阵中，不加批判地选出“最好”的、也就是在假设验证中最显著的相关系数，而没有意识到由于一型差误累积而随机产生的相关性，以及由此产生的风险，即被人为因素欺骗。

1.7.1 产生错觉相关和偏相关

在关联计算中，人们经常根据得出的数值可以推断出具有大甚至很大的相关性。例如，居民区品质和癌症死亡个案之间的相关，未成年人年龄和体质指数之间的相关，或者送子仙鹤数量和每年出生率之间的相关。

由于缺乏解释，这些口口相传的相关性还在盛行。只有通过认证的内容检验和统计检验，才能对这种所谓的错觉相关进行界定。例如，送子仙鹤和每年新生儿数量之间的相关性。送子仙鹤这个个案根本经不起对其内容的检验，因此完全不需要统计检验，而另外两个个案的情况有所不同。

对于居民区品质（通过每平方米价格体现）和癌症死亡个案之间的关联而言，忽略了居民年龄在各个居民区中的不平均分布。只有夫妻两人都工作的年轻家庭才住得起价格昂贵的居民区，相反老年人只能居住在条件简陋的居民区。这些居民区的居民癌症死亡率的高低是和居民年龄

有关，而不是和居民区有关。因此，实际上居民区和癌症死亡率之间的关联是一种（开始没有发现的）产生错觉相关。

利用偏相关的方法，可以根据两个变量之间的关联计算出第三变量可能的影响。在偏相关中，根据 QMPREIS（居民区每平方米价格）和 KREBSTOD（癌症死亡）两个变量计算出变异分量，根据 ALTER（年龄）变量中的变异可以线性地对这个变异分量做出预测。然后将 QMPREIS 和 KREBSTOD 残差相互相关（在一般的半偏相关中，只根据 KREBSTOD 计算出 ALTER 变量的变异分量；然后将 KREBSTOD 的残差与 QMPREIS 相关）。

<pre>CORRELATIONS /VARIABLES= QMPREIS WITH KREBSTOD /PRINT=TWOTAIL NOSIG /MISSING=LISTWISE .</pre>	相关性		
	居民区每平	皮尔逊相关	0.796**
	米价格	显著性（2 位数）	0.000
		N	344
**相关性在 0.01（2 位数）的水平上是显著的。			

```
PARTIAL CORR
/VARIABLES=
  QMPREIS KREBSTOD BY ALTER
/SIGNIFICANCE=TWOTAIL
/MISSING=LISTWISE .
```

相关性				
控制变量			居民区每平方米价格	癌症死亡率
年龄	居民区每平方米价格	相关性	1.000	0.351
		显著性（2 位数）	.	0.000
		自由度	0	341
癌症死亡率	癌症死亡率	相关性	0.351	1.000
		显著性（2 位数）	0.000	.
		自由度	341	0

CORRELATIONS 部分测定了居民区（变量 QMPREIS）和癌症死亡率（变量 KREBSTOD）之间的关联；该关联是显著的、高数值的正相关（ $r=0.796$ ， $p=0.000$ ）。为了简便起见，在本例中默认为实施相关分析的其他条件已经得到满足。

PARTIAL CORR 部分根据 QMPREIS 和 KREBSTOD 之间的关联计算出变量 ALTER 可能的影响。结果十分明显：居民区和癌症死亡率之间的关联尽管是显著的，但是很小（ $r=0.351$ ， $p=0.000$ ）。目前，显著性也受到样本量（ $N=344$ ）的影响。PARTIAL CORR 是基于皮尔逊相关系数的，就这方面而言必须满足实施相关分析的所有条件。在本例中，默认为所有变量的前提条件都已满足（重要的是，应注意个案或者缺失值的数量基本相同）。

未成年人年龄和体质指数之间的全面关联尽管是不显著的，但仍有很大的信息量。下面这个例子涉及的不是一个定距变量，而是一个分类干扰变量。分类干扰变量的效应无法通过偏相

关于以检验，但是可以首先简单地通过“组成组块”进行相互比较。

```
CORRELATIONS
/VARIABLES=
  alter WITH bmi
/PRINT=TWOTAIL NOSIG
/MISSING=LISTWISE .
```

相关性		
		BMI
年龄	皮尔逊相关	0.231**
	显著性 (2 位数)	0.005
	N	144

**相关性在 0.01 (2 位数) 的水平上是显著的。

```
SORT CASES BY gschlcht .
SPLIT FILE
  LAYERED BY gschlcht .
```

```
CORRELATIONS
/VARIABLES=age WITH bmi
/PRINT=TWOTAIL NOSIG
/MISSING=LISTWISE .
```

```
SPLIT FILE
  OFF .
```

相关性		
性别		BMI
男性	年龄 皮尔逊相关	-0.028**
	显著性 (2 位数)	0.812
	N	77
女性	年龄 皮尔逊相关	0.502**
	显著性 (2 位数)	0.000
	N	67

**相关性在 0.01 (2 位数) 的水平上是显著的。

如果将数据按照人的性别（变量 **GSCHLCHT**）分组，则表现出完全相反的效应。未成年人的年龄和体质指数之间的关联（ $N=144$ ）尽管是显著的，但是很小（0.231），而且针对男孩和女孩分别测定的相关会得出完全不同的结果。对于男孩（ $N=67$ ），年龄和体质指数之间的关联根本没有呈现出显著性；而对于女孩（ $N=77$ ），年龄和体质指数之间的关联不仅达到显著性（ $p=0.000$ ），而且程度相当可观（0.502）。因此，取得的结果十分复杂。

开始测定的相关性可能也同时受到了样本量（ $N=144$ ）的影响。此外，第一个结果的非特定性起到了误导作用，使人相信变量“性别”可能有差异化的效应，这个效应表现为年龄和体质指数之间的关联具有性别上的差异。在相互比较的群组中，个案或者缺失值的数量基本相同并且也很重要。后面将介绍如何直接比较不同的相关系数。对于定类型数据，在建立交叉列表（如 **CROSSTABS**）时还可以通过引入控制变量来计算偏相关。

因此从统计角度来看，产生错觉相关的原因是由于两个被检验变量之间的第三变量制造出虚假关联而出现的，如果解释时没有考虑到第三变量的影响，则会产生解释错误的危险（参见 Cohen 等人，2003，75-79，Litz，2000，77-91；Pedhazur，1982，110-111）。

在通过偏相关计算出干扰效应时，应只选择从逻辑上予以考虑的或者确实存在的变量。如果没有调查这些变量，则无法排除其效应。

在计算偏相关时，应清楚地了解相关理论，以及用这种方法可以调查的因果模型。例如，根据 Pedhazur 的著作（1982，110-111），可以选用下列两个模型结构。

$X \rightarrow Y \rightarrow Z$ 例如， $QMPREISE \rightarrow ALTER \rightarrow KREBSTOD$	$X \leftarrow Y \rightarrow Z$ 例如， $QMPREISE \leftarrow ALTER \rightarrow KREBSTOD$
--	--

在第一个模型中, Y 影响 X 到 Z 的效应 (Y 是中介变量); 在第二个模型中, Y 是 X 和 Z 的共同原因, 造成了 X 和 Z 之间产生的错觉相关。

当要调查交互作用或者间接效应 (中介变量) 时, 如果存在多个关系作为预测变量, 例如, 其中 X 和 Z 既没有直接地相互影响, 也没有间接地通过 Y 相互影响, 或者 X 和 Y 是 Z 的相关原因时, 就不适合使用因果相关了。对于这些模型和其他模型, 路径分析提供了合适的分析技术。

1.7.2 一型误差累积问题

即使有些作者是在描述性统计量的篇章中介绍相关分析, 但他们在从样本到总体的推导时还是以假设检验为依据。将假设表述出来, 进行检验, 然后再接受或者拒绝这个假设。在这里需要注意的是, 相关是基于两个变量的。但是, 通过计算机辅助数据分析 (不管是通过界面中选择选项还是编写语句) 时, 人们可以调用远远超过两个变量的相关, 然后从大量显著的相关中找出令人印象最为深刻的相关。问题在哪里? 这种“寻找显著性”的处理方法有很多问题, 作为依据的“追求显著性”就有问题, 就好像令人印象最深刻的关联比其他关联更重要一样。事实完全相反, 这两种调查结果从形式上来看是同等重要的。但是, 这种显著性固定现象除了显著性结果之外, 还会导致严重的“发布错误”。无法展现这方面区别 (或者是由于根本不存在区别) 的论文, 早先由于“追求显著性”和一型误差累积而没有得到发布 (例如, Witte, 1980, 51-59; Bredenkamp, 1972, 53), 从而导致对现实情况 (如临床现实) 的表现偏误 (例如, Turner 等人, 2008; Hackbarth, 2008)。在出版的杂志, 例如, 生物医学负面研究成果杂志中提出了一条具体的针对性措施, 其目的是仅仅发布不显著的结果。

下一个问题是, 这种分析实践不再符合传统的假设检验; 也就是没有表述任何假设, 而是一次性分析所选择的所有关联, 查找显著性, 然后不加考虑地将其发布 (这种处理方法很可能没有检验是否满足了调用相关系数的所有前提条件)。这是典型的“寻找显著性”处理方法, 即假设实施的相关分析越多, 找到显著性的概率就越大。但是这没有考虑到, 由于实施了大量相关分析 (这一点原则上适用于所有重复使用的多变量方法), 有可能随机出现显著性。根据经验法则, 连续实施 100 次检验就会随机出现 5 次显著性。因此, 这种“寻找显著性”处理方法的不科学之处在于, 它不仅对随机出现的显著性一无所知, 而且无法将其与根据假设而推导出的显著性 (或非显著性) 区分开来。

设想一下, 有个飞镖游戏者闭着眼睛将飞镖投向标靶 100 次, 每次投掷就相当于实施一次相关分析, 每投中一次标靶就看作是出现一次显著性。但是, 这样的投中结果既不能证明游戏者飞镖玩得好, 也不能说明应对其予以认真对待, 因为它是没有经过检验或者理论推导就产生的。现在想象一下, 这个飞镖游戏者把他闭着眼睛投中 5 次标靶的结果公诸于众, 而没有说明他进行了多少次检验 (投掷) 才得到这个结果。分析 (投掷) 实施的次数越多, 得出的显著性结果 (投中) 就越夸大、不可信 (因为是随机的)。无论是对于科学研究还是飞镖游戏 (这也可以是一种科学), 这样的处理方法都是很可疑的。

根据假设检验的典型图形可以知道, 正确地接受零假设的概率是 $(1-0.05) = 0.95$ 。0.05 在这里指的是一型误差, 也就是在尽管零假设确实正确的情况下接受显著性的概率。如果进行两次 (独立) 检验, 则这个概率更低: $0.95^2 = 0.90$, 三次检验得到 $0.95^3 = 0.86$, 四次是

$0.95^4=0.81$ ，五次是 $0.95^5=0.77$ 。相反，一型差误从最初的 0.05 明显升高到 0.10、0.19，直至 0.23。这个现象通常称为一型差误累积，指的是随着检验次数的增加，将随机出现的显著性误认为正确而予以接受的概率也随着升高。

示例

如果只有 5 个变量相互相关，则产生 10 次不同的假设检验；如果觉得公式 $(N/2) * (N-1)$ [在本例中 $N=5$] 太抽象，则可以将下表对角线左栏或者右栏的数值倒着数。正确地接受一个零假设的概率只有 0.60；相反，错误地接受一个零假设的概率达到 0.40。这样就差不多成了纯粹的猜测。另一个问题是：利用 SPSS 测定的显著性没有针对一型差误累积而做调整。用户必须自己进行这样的调整。

但是，可以相对简单地用 Bonferonni 校正对一型差误累积采取对策。在进行 Bonferroni 校正时，只需简单地将测定的显著性值与假设检验次数（在这里是 10 次）相乘。如果乘得的显著性值低于 α 初始值（在这里是 0.05），则可以将其视为显著。

相关性					
	kör	ene	psy	soz	ess
kör 皮尔逊相关	1	0.613**	0.517**	0.454**	0.210**
显著性（2 位数）		0.000	0.000	0.000	0.003
N	191	191	191	191	191
ene 皮尔逊相关	0.613*	1	0.572**	0.620**	0.144*
显著性（2 位数）	0.000		0.000	0.000	0.046
N	191	191	191	191	191
psy 皮尔逊相关	0.517**	0.572**	1	0.418**	0.092
显著性（2 位数）	0.000	0.000		0.000	0.208
N	191	191	191	191	191
soz 皮尔逊相关	0.454**	0.620**	0.418**	1	0.197**
显著性（2 位数）	0.000	0.000	0.000		0.006
N	191	191	191	191	191
ess 皮尔逊相关	0.210**	0.144*	0.092	0.197**	1
显著性（2 位数）	0.003	0.046	0.208	0.006	
N	191	191	191	191	191

**相关性在 0.01（2 位数）的水平上是显著的。

*相关性在 0.05（2 位数）的水平上是显著的。

在上表中，ESS 和 ENE 之间关联的 $p=0.046$ ，因此是（有可能随机）显著的。在根据假设检验的实施次数进行校正后，得到 $p=0.46$ ，因此很明显不再是显著的。ESS 和 KÖR 之间的关联在经过 Bonferonni 校正后仍是显著的（未调整：0.003；校正后：0.03）。

1.8 特殊用途

本节将介绍如何直接比较相关系数、检验相关的一致性以及计算正准相关。

1.8.1 相关系数的比较

SPSS 可以通过几种方式进行相关系数的比较。第一种方式是，在针对总体（POPCOEFF）的相关系数说明样本量（NSAMPLE）时，检验样本（CSAMPLE）的相关系数。这种方法两个假设是，总体的相关系数不等于零，并且作为依据的 z 值基本上是标准正态分布的。零假设的含义就是：样本的相关系数等于总体的相关系数。

在男孩和女孩的体质指数与年龄之间具有不同相关性的例子中，分别将相关系数用作基本数据。男孩和女孩（ $N=144$ ）的年龄与体质指数之间的相关性总共达到显著的 0.231；在本例中，将其作为总体的相关系数引入分析。但是，男孩和女孩在相关上的表现完全不同。在男孩（ $N=67$ ）中，年龄和体质指数之间的相关性（-0.028）完全达不到显著性。对于女孩（ $N=77$ ），同样的相关性则达到了显著的 0.502。

在下面的 COMPUTE 例子中，针对未成年人相关系数的总值（0.231， $N=144$ ），检验女孩的相关系数（0.502， $N=77$ ），下一个步骤则用同样方式检验男孩的相关系数。每次只将四舍五入的数值引入分析。实施的检验是双侧的，但是总体的相关系数不得为零。在 COMPUTE 例子中，也可以给出几个负值的相关系数。但是 POPCOEFF 值不得为零。

```
data list free
/ CSAMPLE NSAMPLE POPCOEFF.
begin data
0.50 77 0.23
-0.03 67 0.23
end data.
compute #ZSAMPLE = .5* (ln ( (1 + CSAMPLE) / (1 - CSAMPLE) ) ) .
compute #ZPOP = .5* (ln ( (1 + POPCOEFF) / (1 - POPCOEFF) ) ) .
compute Z = (#ZSAMPLE-#ZPOP) / (1/ (sqrt (NSAMPLE-3) ) ) .
compute PWERT = 2* (1-cdfnorm (abs (Z) ) ) .
format PWERT (F8.3) .
list.
```

根据 DATA LIST 将各个样本（女孩，男孩）的相关系数读取为 CSAMPLE，根据 NSAMPLE 读取样本的大小，并将总体的系数读取为 POPCOEFF。

输出结果

CSAMPLE	NSAMPLE	POPCOEFF	Z	PWERT
0.50	77.00	0.23	2.71	0.007
-0.03	67.00	0.23	-2.11	0.035

```
Number of cases read: 1      Number of cases listed: 1
```

每次测定的显著性（PWERT）低于 0.05 的 α 值。女孩的相关系数从统计学角度来看与总体的相关系数（ $p=0.007$ ）有显著区别，男孩的相关系数也是如此（ $p=0.035$ ）。

SPSS Command Syntax Reference 含有用于这种检验的一个宏，但是这个宏不能给出样本的负值相关系数。

对相关系数的比较，只有在考虑到作为依据的极差或者在 X 、 Y 轴上的分布位置时才有意义（参见第 1.5.2 节）。

1.8.2 比较相关的一致性

如果有分组的数据，例如，根据性别分组的年龄和体质指数，则可以检验不同组之间是否有同样双变量相关的零假设。在本例中，就是检验男孩和女孩体质指数与年龄之间的相关是否等同。这种解决方法的依据就是，根据 z 标准化变量检验斜率（回归同质性）的一致性，此时回归斜率等于相关性。然后将这些变量引入协方差分析，此时对分组变量（在这里是 **GSCHLCHT**）和两个 z 标准化变量中的一个变量（是哪一个并不重要）之间的交互作用来检验，不同组的斜率（相关性）相等这个零假设是否成立。

```
DESCRIPTIVES
  VARIABLES=age bmi
  /SAVE
  /STATISTICS=MEAN STDDEV MIN MAX .

UNIANOVA
  zBMI BY gschlcht WITH zage
  /METHOD = SSTYPE(3)
  /INTERCEPT = INCLUDE
  /CRITERIA = ALPHA(.05)
  /DESIGN = gschlcht*zage .
```

在利用 **DESCRIPTIVES** 测定 z 标准化变量时应注意，个案或者缺失值的数量与所实施相关或者协方差分析的次数一致。

组间因子			
		数值标签	N
性别	1.00	男性	77
	2.00	女性	67

组间效应检验

因变量 Z 值: BMI					
来源	三型平方和	df	平方均值	F	显著性
调整模型	19.953 ^a	2	9.977	9.671	0.000
常量项	17.070	1	17.070	16.548	0.000
性别*ZAGE	19.953	2	9.977	9.671	0.000
错误	145.451	141	1.032		
总和	180.591	144			
调整的总变化	165.405	143			

a. $R^2=0.121$ （调整 $R^2=0.108$ ）

分组变量（在这里是性别）和 z 标准化变量 **ZAGE**（年龄的 z 标准化）之间的这种交互作用是显著的。拒绝了斜率相等（相关性）这个零假设。男孩和女孩的斜率（相关）是不同的。

但是，在这里比较相关性时没有考虑到作为依据的极差或者在 X 、 Y 轴上的分布位置，因此意义有限（参见第 1.5.2 节）。

1.8.3 正准相关

利用正准相关（又称 set correlation）可以测定两组（集合）变量之间的线性关联，例如，一组预测变量和一组准则。因此，当调查两个分别由几个变量组成的特征的关联时，使用这种方法就非常有效。正准相关方法的最大问题在于，所测定正准相关系数的可解释性。最大的线性相关有时会造成可解释性最小。

正准相关方法简述

在使用正准相关方法时，在两组中的每一组中，首先测得具有目标特性的变量（即正准变量）的线性组合，两个正准变量 K_1 和 K_2 之间的相关达到最大。 K_1 和 K_2 之间的相关是第一个正准相关。在提取了 K_1 和 K_2 后，针对每个集合都保留了一个剩余方差，对其同样测定相互相关性最大的线性组合（测定的正准相关逐渐变小）。不断重复这个过程，直到总方差在两个集合中的一个中穷尽为止。正准相关的数量绝不等于两个集合的任何一个的变量数量。正准相关的数值始终大于多重相关的最大值。一个集合中的正准变量相互相关性为零（正交性）。

示例

```
GET FILE "C:\...\IHREDATEN.SAV".
INCLUDE "C:\...\SPSS\Canonical correlation.sps".
CANCORR
    SET1=ENE ESS
    / SET2=KÖR PSY SOZ.
```

正准相关的计算并不复杂。通过 GET FILE 命令首先调用一个数据集，要引入的正准相关的变量就在这个数据集中。通过 INCLUDE 命令将“Canonical correlation” SPSS 宏引入分析。原则上无须对宏进行调节，通过 INCLUDE 命令引入就足够了。这个宏在大多数情况下位于子目录“.../SPSS”中。利用 CANCORR 过程，只给定两个集合 SET1= 和 SET2= 的变量（在本例中就是第一个集合的变量 ENE 和 ESS，第二个集合的 KÖR、PSY 和 SOZ），针对这些变量测定正准变量 K_1 至 K_n 或者相应的正准相关。

输出结果（节选）

Correlations Between Set-1 and Set-2

	kör	psy	soz
ene	0.6132	0.5715	0.6201
ess	0.2104	0.0916	0.1974

Canonical Correlations

1	0.761
2	0.115

Test that remaining correlations are zero:

Wilk's	Chi-SQ	DF	Sig.
--------	--------	----	------

1	0.415	164,596	6,000	0.000
2	0.987	2,489	2,000	0.288

在“Correlations Between Set-1 and Set-2”部分首先输出多重相关系数。在“Canonical Correlations”部分，输出两个集合 SET1 和 SET2 之间的正准相关系数，其值为 0.761 和 0.115。在“Test that remaining correlations are zero”下，根据每次测定的正准相关检验零假设，即假设下一个正准相关等于零。

根据第一个正准相关，拒绝了零假设（ $p=0.000$ ），实际上还测定了第二个正准相关，其值为 0.115。但是，根据第二个正准相关，保留零假设（ $p=0.288$ ），不再继续测定正准相关。

这种方法最重要的前提条件是，相关-协方差或者方差-协方差矩阵是建立在各个成对变量之间线性关联的基础上的。此外，变量应是定距的、多变量、正态分布且没有离群值。对于二元预测变量，准则变量必须在所有由预测变量定义的子总体中呈正态分布。如果不存在线性关联，则可以转为利用 SPSS 过程命令 OVERALS 进行非线性的正准相关分析（菜单项：“最佳尺度...”）。

1.9 计算皮尔逊相关系数的前提条件

为了计算皮尔逊相关系数，应注意很多前提条件或者特别之处，下面将予以综述（参见 Cohen 等人著作，2003）。

1. 相关无法检验两个变量之间的因果关系。如果不存在双变量相关，则没有双变量因果关系。
2. 从一开始就应根据逻辑排除产生错觉相关。例如，送子仙鹤数量和新生儿数量之间的相关。
3. 测量值对 x_i 和 y_i 必须属于同一个对象。换言之，所调查的特征是从一个样本的同一个元素中提取的。
4. 对于所需的样本量范围有几个经验法则，即建议最少 $N=50$ （Green，1991）或者 $N=30$ （Borg & Gall，1989）。如果呈现偏态分布或者数据含有测量误差，则需要更多的个案。如果数据近乎完美地线性排列，并且要计算理想关联的相关性，则允许使用较少的成对测量值（例如， $N=6$ ）。
5. 变量都是定距的。散点图中的阶梯形图表明，两个变量中至少有一个不是定距的。
6. 测量误差（可靠性，偏倚）：两个变量应是十分可靠的，也就是在理想情况下测得而且没有误差。测量误差将减小相关性。双变量偏倚将使相关性过高（例如，选择机制，将两个变量除以同一个分子或者分母，或者在各项与复合量度相关时，即部分-整体相关）。
7. 如果要从一个样本推断出总体，例如，通过显著性检验，则两个变量的数据必须分别呈现正态分布。如果没有呈现二维正态分布，则无法计算非线性关联或者在较低尺度水平上的关联。

8. 在两个变量之间存在线性关联。如果两个随机变量之间的关联是非线性或者曲线形的，则皮尔逊相关系数不适合用于计算关联的程度。如果成对测量值基本上相互平行（尽量）地散布在相关直线两侧，则具有同方差性（散布的稳定性）。
9. 相反，如果有异方差性，则测定作为平均值的标准估计误差低估了宽松的散布状况，并且高估了紧凑的散布状况，这样就不再适合用作描述线性关联的平均值。在这种情况下，可以通过组建子群将总体散布分开，针对各个分组分别测定相关系数。断续直线的类似现象需要用另一种处理方式。
10. 相关直线应具有连续性。如果一条直线有（一个或者几个）间断，则可以将其解释为样本无代表性。应通过继续取样填补样本，使间断得到闭合。否则就会导致一个双重问题：无代表性样本产生偏误的估计。
11. 没有离群值。应删除基于扰动影响的离群值，因为离群值会导致产生误导性的相关（尤其是样本较小时）。朝着函数方向的离群值使相关系数过高；远离于函数的离群值导致对相关系数的估计过低。来自于无代表性的样本会导致产生离群值的间断很大，此时应通过继续取样予以填补。
12. 相关系数的大小也取决于两个变量的变异性。如果忽视了两个变量中至少一个的极差，则就人为地降低了相关系数（参见 Cohen 等人著作，2003，57）。只有通过有代表性的样本才能解决这个问题。相关性应既没有超出现有的测量值域，也不应在测量值域之外被解释。
13. 显著性同时也受基础样本量的影响。显著性并不总是意味着关联性。
14. 第三变量的效应：相关可能受其他变量的影响。子群的相关系数可能与总体的相关系数有所区别。根据调查的对象不同，这种效应可能完全不同。因此，在任何情况下应保护相关系数不受异质子群的人为影响。产生错觉相关，即不是由两个受调查的变量，而是由第三变量造成的相关，它可以从概念和计算的角度，也就是通过计算半相关或偏相关而予以删除。
15. 在同一方向上的变量是极化的（尤其是在解释内相关的心理测量尺度时很有帮助）。
16. 在解释相关系数时，应注意期望关联和检验的方向。从显著性方面来看，单侧检验的结果就已经无法接受了；在检验前有约束力的、确定的关联（正关联或者负关联）必须已经出现。

第 2 章 线性回归和非线性回归

第 2 章介绍回归分析。本章内容采用逐步推进的架构，以阐明实行回归分析时的基本原则，并帮助读者从一开始就避免常犯的错误。

第 2.1 节首先介绍了简单线性回归分析（SPSS 过程命令 **REGRESSION**）。第 2.1 节通过一个简单的例子，阐述了如何根据杠杆值和残差来检验线性和识别离群值。还阐述了如何检验可能存在的自相关。一般来说，利用线性回归分析只能调查线性函数。利用线性回归分析来调查非线性函数会产生错误的结果。不是所有函数都能满足简单线性回归的前提条件。因此第 2 章将尽可能简单地从线性回归过渡到非线性回归，并帮助解答下列问题：如果两个变量之间的关联是非线性的怎么办？如果双变量线性回归模型的残差不是随机分布，而是非线性分布的怎么办？如果在一个双变量线性回归模型中存在异方差性怎么办？

第 2.2 节阐述了如果数据不是线性而是曲线分布时应该怎么做。第 2.2 节提供了两种解决方案：将非线性函数进行线性化，并用线性回归进行分析。也可以用非线性回归对非线性函数进行估计（SPSS 过程命令 **CNLR** 和 **NLR**）。非线性回归是本节的中心主题，包括带有两个预测变量的非线性回归。此外，阐述了用于（非）线性曲线估计的 SPSS 过程命令 **CURVEFIT** 的意义和限制。最后几段总结了非线性回归的各种假设，并通过一个总览表介绍了最有名的一些非线性回归模型，其中含有一个或多个预测变量。

第 2.3 节介绍了多元线性回归分析（SPSS 过程命令 **REGRESSION**），与简单线性回归相反，这种分析较为复杂。模型含有多个，而非仅有一个自变量时，就会使模型主要是在自变量相互之间关系方面具有一些特点。本节着重探讨了建模、变量选择、多重共线性和其他难点。除了识别和消除多重共线性外，也探讨了如何处理时间相依（自回归）数据。通过第一个例子，

介绍了多元回归中之重重中之重重中之重重中之重作者在归的特性参数；通过第二个例子论述了多重共线性的问题，以及如何识别和处理多元共线性。并且对偏回归的计算也给出了说明。最后一段归纳了实施（非）线性回归分析的各种前提条件以及对其进行检验的方法（见第2.4节）。

2.1 线性回归：有因果方向的关联

本节的开始部分概述了简单线性回归（参数的），以及如何在 SPSS 中利用鼠标简单地进行调用。根据输出的语句，示范性地检验已实施分析的给定选项是否恰当（参见 Schendera, 2005）。如果确保了计算正确进行，则首先阐述图形式残差分析中很有用的图形，然后解释所调用的回归分析的统计量。

2.1.1 双变量线性回归：利用 REGRESSION 的回归分析概述

类似于相关分析，通过回归分析也可研究两个变量之间是否存在关联。除了计算方法本身之外，还存在至少三个根本性区别（见下文）：回归分析假设了一个因果模型（如“ X 导致 Y ”），从而不仅说明了两个变量之间是否以及在多大程度上存在关联，而且还检验了关联的方向，即因果模型本身的方向，也就是自变量（UV、回归量、预测变量、解释变量、影响变量）在多大程度上影响因变量（AV、回归应变数、标准、目标变量）。

在回归分析中所使用的术语，有时区别在于研究目的究竟是预测（预测性）还是解释（解释性）。产生这种区别的主要原因是，回归分析常常可以做出预测，但不是总能同时做出解释。预测性回归分析往往只使用“回归量”/“回归应变数”或是“预测变量”/“标准”这样的成对概念，而解释性回归分析往往只使用术语“解释变量/效应”（Pedhazur, 1982², 135-137）。

下列问题是回归分析的应用示例：

- 所需劳动力的数量对其薪酬有影响吗？
- 妊娠期对新生儿体重有影响吗？
- 发动机功率对汽车耗油量有影响吗？

如果把这些问题的表达与相关分析的示例进行比较，就会确定这里设定了一个因果方向。除此之外，根据所猜测影响的（正、负）方向和程度大小也能区分这些问题。

区别于相关分析，回归分析具有给定的因果方向。在进行回归分析之前，从理论上确定影响方向（确定变量是自变量还是因变量）。此外，前提条件是自变量和因变量之间存在线性关系。与相关分析所提出的问题（即“一个变量被另一个变量的影响程度有多大，或者反而言之？”）相反，回归分析所提出的是“当自变量发生有规律的变化时，因变量将如何变化？”。与此关联的是，第二个区别涉及自变量或者因变量的成立；这两个变量并不一定是随机变量。例如，自变量值经常发生有规律的变化，而因变量的值相应地也发生变化，因此因变量不一定是随机变量（Pedhazur, 1982², 33-40）。第三个区别涉及变量的数量。相关总是仅仅由两个变量的关联组成，而相关系数矩阵则由许多双变量的关联组成。但是，回归分析可以从一个简单的双变量回归扩展成多变量回归，即所谓的多元回归（参见第2.3节）。

在双变量个案（简单线性回归）中，从自变量 X 推导出因变量 Y 。在多元回归中，通过多

个自变量 X_1, X_2, \dots 的线性组合预测因变量 Y 的值。

然而这仅仅涉及自变量的数量。利用回归的标准模型，通常只能计算出对单个变量的影响。如果要检验一个模型，并且这个模型同时描述了对多个因变量的影响，那么就应选择路径分析作为检验方法（参见第 6 章关于 ANOVA/MANOVA 的内容）。

例如，当所声称的因果性表现出“显著性”时，应谨慎地看待回归分析的“证据”（也与显著性的绝对化相关，参见 Witte, 1980）。这样，回归分析不排除其他竞争（如表述完全相反的）模型的有效性。但是如果没有达到显著性，也就不存在因果性，至少不存在所断定形式的因果性。

除了简单的（双变量线性）回归（过程 REGRESSION，也称为过程 GLM）外，SPSS 还提供其他的回归：

- 逐步回归（REGRESSION，选项 STEPWISE）。
- 向前回归或者向后回归（REGRESSION，选项 FORWARD 或 BACKWARD）。
- 非线性回归（SPSS 过程命令 NLNR，CNLG）。
- 多元线性回归（REGRESSION，也称为过程 GLM）。
- 分类回归（过程 CATREG，LOGISTIC REGRESSION，PLUM，NOMREG 和 PROBIT）。
- Cox 回归（SPSS 过程命令 COXREG）。
- 偏回归（SPSS 过程命令 PLS）。
- 加权最小二乘回归（SPSS 过程命令 WLS）。
- 2SLS 回归（2 级 LS 回归，SPSS 过程命令 2SLS）。
- 岭回归（SPSS 宏“Ridge-Regression（岭回归）宏”）。
- 用于曲线估计的过程 CURVEFIT。
- 也可以分别进行多元线性回归，从而计算出偏回归。

下面阐述简单线性回归的基本思想和计算步骤。

推荐 Chatterjee & Price (1995²) 和 Pedhazur (1982²) 的著作作为回归分析的入门文献。

简单线性回归的基本思想

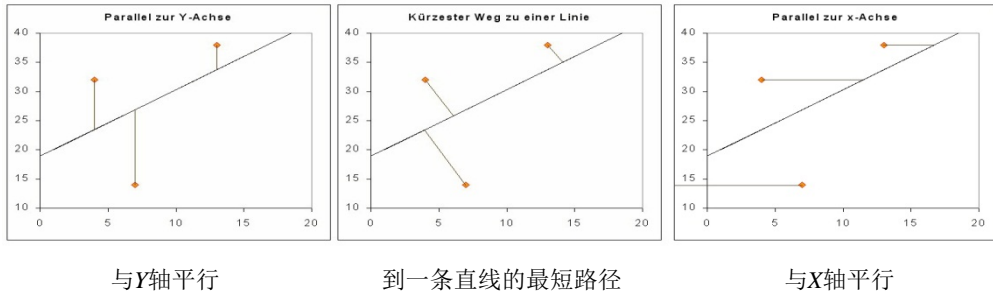
只有当两个变量 X 和 Y 之间存在（尽可能高的）关联，计算线性回归才是有意义的（需要注意的是，如果关联较低则可能存在曲线关联，此时非线性回归可能更为适合）。这些成对测量值在散点图（也就是 X - Y 坐标系）中构成了理想的、线状或带状的点排列（参见本章中关于散点图的例子）。如果这些点呈带状散布在假想线上面或者周围，那么对于线性走向的假设就是成立的，通过线性函数就可以描述这个散点图。之所以是线性函数，是因为回归函数的图形为一条直线。相反，如果是曲线状的散点图，则非线性（如二次回归或三次回归）回归函数是适当的。有些函数也可以通过转换实现线性化。

确定回归直线

从计算方面，人们用最小二乘法（又称 OLS，Ordinary Least Squares “普通最小二乘法”；不要与 WLS，Weighted Least Squares “加权最小二乘法”混淆）确定穿过散点图的直线。最小

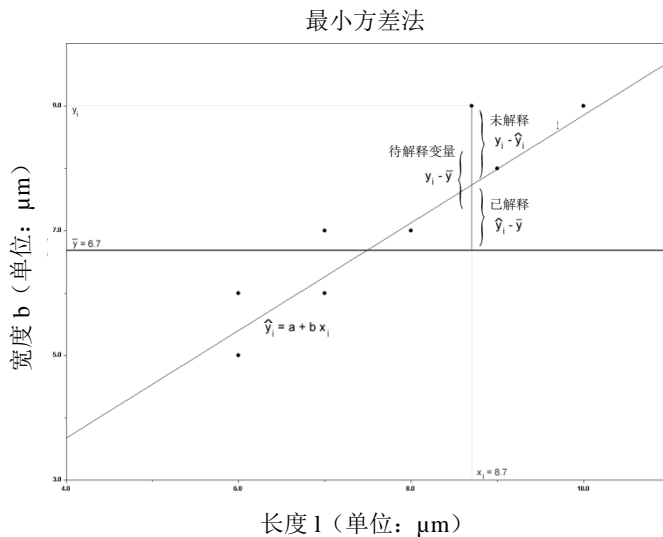
二乘法将观察值与朝着变量 Y 方向的回归直线之间的距离最小化，从而使所有的点与待定直线之间距离的平方和最小。因此，用最小二乘法计算得出的直线通常是描述变量 X 和 Y 之间线性关联的最佳估计直线。

通常，人们把两点之间的最短路径称为距离。一个点 (x_i, y_i) 与一条线之间的最短距离是指从该点出发，并垂直于这条线的连接线的长度（见中间的图示）。但对于回归而言，距离的测定总是沿着因变量的方向进行的；对此的约定是，因变量是由 Y 轴表示的。因此标尺并不垂直于直线，而是平行于 Y 轴。（见左图，也可参见 Pedhazur 的著作，1982²，22-24。）



图：点到直线距离的测定。左图：平行于 Y 轴，中图：点到直线的最短路径，右图：平行于 X 轴。右图利用相同的数值展示了如果平行于 X 轴测量距离，将得到完全不同的结果。

如果从图中 Y 轴上任意一段截取所估计的平均值，再画一条平行于 X 轴、延伸到估计函数并且长度为这个平均值的直线，最后从任意一点 (x_i, y_i) 画一条平行于 Y 轴、延伸至函数 ($\hat{y} = a + bx$) 的直线，就会得出三个距离值。



点 (x_i, y_i) 到 \bar{y} 的距离表示待解释离差，可以分解为解释离差 $(\hat{y}_i - \bar{y})$ 和未解释离差 $(y_i - \hat{y}_i)$ 两个分量。因此，回归分析的目的就是将可解释离差 $(y - \bar{y})$ 分解为被解释离差 $(\hat{y}_i - \bar{y})$ 尽可能大的分量和未解释离差 $(y_i - \hat{y}_i)$ 尽可能小的分量： $(y - \bar{y}) = (y - \hat{y}) + (\hat{y} - \bar{y})$ ，并使离差的平方和最小： $\sum (y_i - \hat{y}_i)^2 = \text{Minimum}$ 。

对回归函数的解释

回归直线基于方程 $y = a + b \cdot x$ 。所以回归直线不仅可以描述现有数据，而且基于这个方程可以做出统计解释。

因此，回归方程可以用于描述、估计、预测及建模。如果已知 a （截距，常量）和 b （斜率），就可以知道相关直线的准确位置。根据所了解到的成对测量值中的一个变量 X 的值，回归函数可同时描述、估计并预测这个成对测量值中另一个变量 Y 的值。

如果用 y_i 来表示对于 x_i 的测量值，则用 \hat{y}_i （ y 的估计值）表示在相应回归直线上的 y 值。所谓的最小属性是指：各点与某条直线的距离平方和小于与其他任何直线的距离平方和；从数学上表达就是： $\sum_{i=1}^n (\hat{y}_i - y_i)^2 = \text{最小}$ 。回归直线是一条直线，因此 $\hat{y}_i = a_x + b_x \cdot x_i$ 成立。把这个等式代入第一个方程中，会得出 $\sum_{i=1}^n (a_x + b_x \cdot x_i - y_i)^2 = \text{最小}$ 。利用微积分知识，针对 x 对于 y 的回归直线（以及用于演示 y 对于 x 的回归，见下文）可以得出：

x 对于 y 的回归

$$b_x = \frac{\sum_{i=1}^n ((x_i - \bar{x}) \cdot (y_i - \bar{y}))}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$a_x = \bar{y} - b_x \cdot \bar{x}$$

y 对于 x 的回归

$$b_y = \frac{\sum_{i=1}^n ((x_i - \bar{x}) \cdot (y_i - \bar{y}))}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$$a_y = \bar{x} - b_y \cdot \bar{y}$$

观察值和由回归方程得出的估计（预测）值之间的（理想情况下应尽可能地小）差值被称为残差（又称残数、误差）。如果观察值（理想情况下精确地）位于所画出的回归直线上，则残差等于零（观察值和等于相应的估计值）。残差的总离差越小，回归函数对点散布情况的解释就越准确。如果这个模型的所有值都准确地位于这条回归直线上，那么这个回归方程就完美地描述了变量之间的关联。在这种条件下， R 和 R^2 会达到可能的最大值。离差越大，这个回归方程的可信性就越小。为了评估这个模型，残差和预测值还必须满足其他特性，尤其是正态分布特性。

有人认为，通过对残差进行仔细的（图形式）分析来评估模型适用性，比计算回归更为重要（如 Chatterjee & Price, 1995²）。

这个说法的正确性得到越来越多人的认可，因为只观察统计特征量很可能产生误导。例如，根据基本数据不同，同样的回归系数可能有一次达到显著性，在其他条件下则不能。主要的统计特征量，例如， R^2 、 F 值以及回归方程的显著性等主要受到抽样大小、所采集数据的值域和数值变化幅度的影响。在散点图中可以比在数据列表中更简单地识别到这些特征（参见 Pedhazur, 1982², 30-32）。尤其是对于 R^2 的解释而言，这种现象意义重大。例如，随着变量数量增多， R^2 增大；而随着 N 增大， R^2 重新降低。因此，在个案很多和很少的模型中，较高的 R^2 值具有完全不同的意义。在简单线性回归的个案中， R^2 等于预测变量与反应变量间皮尔逊相关系数的平方（ $R^2 = r^2_{xy}$ ）。

将给定的 x 值带入回归方程，就可以得到期望的 y 值（例如，模型 1，自变量：长度 → 因变量：宽度）。如果想根据给定的 y 值计算出期望的 x 值（例如，模型 2，自变量：宽度 → 因变量：长度），就不能只是简单地转换模型 1 的直线方程，而是必须计算出模型 2 自己的回归函数。

示例：酵母细胞

为了用显微镜测量酵母细胞的长度和宽度，应进行回归分析计算。为了清晰，这里对两个因果模型进行计算。首先应确定酵母细胞长度对宽度的影响，然后反之确定宽度对长度的影响（参见 Lorenz 在《第二种回归》一书中阐述的概念，1992，72）。

```
DATA LIST FREE
/ LANGE BREITE.
BEGIN DATA
6 5  7 6  6 6  7 6  8 7  9 8  10 9  8 7  7 7  7 6
END DATA.
EXE.
```

由此得出下列回归函数：

x 对于 y 的回归

$$b_x = \frac{\sum_{i=1}^{10} ((x_i - 7.5) \cdot (y_i - 6.7))}{\sum_{i=1}^{10} (x_i - 7.5)^2} = \frac{12.5}{14.5} = 0.862$$

$$a_x = 6.7 - 0.862 \cdot 7.5 = 0.235$$

$$\hat{y}_i = 0.235 + 0.862 \cdot x_i$$

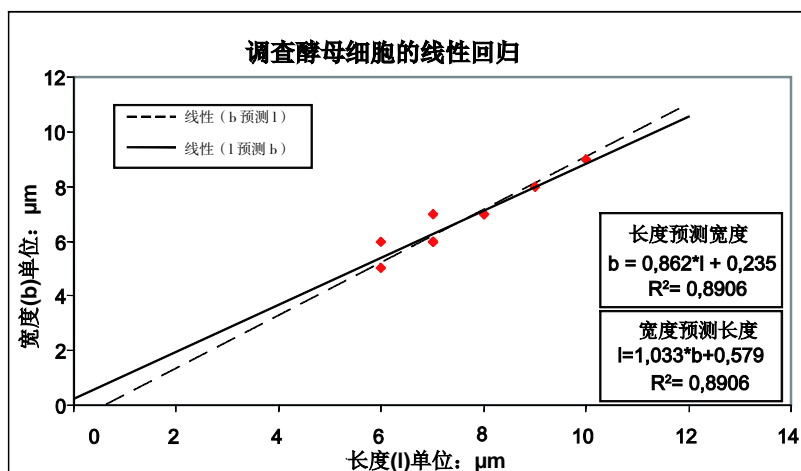
y 对于 x 的回归

$$b_y = \frac{\sum_{i=1}^{10} ((x_i - 7.5) \cdot (y_i - 6.7))}{\sum_{i=1}^{10} (y_i - 6.7)^2} = \frac{12.5}{12.1} = 1.033$$

$$a_y = 7.5 - 1.033 \cdot 6.7 = 0.579$$

$$\hat{x}_i = 0.579 + 1.033 \cdot y_i$$

上述两个回归直线见下图。通常来说，回归不应超出现有的测量值值域。但在这张图中却超出了，其目的是为了清楚地标明回归直线与 X 轴或 Y 轴的交点，即所谓的截距。此外，图中还给出了决定系数，这个系数在两个回归方程中必须是相等的。



这个图形展示了两个模型：实线的回归线表示长度为自变量、宽度为因变量的模型，虚线的直线表示宽度为自变量、长度为因变量的模型的回归函数。在以下内容中，将详细解释图例中给出的特性参数（如系数）以及其他许多特性参数。

2.1.2 双变量线性回归的示例和语句——第一步：根据杠杆值和残差检验线性并识别离群值

在临床数据的基础上，利用 SPSS 对简单（双变量）线性回归进行计算。在本章中，将根据杠杆值和残差进行第一次分析，包括对线性和方差一致性（同方差性）的检验，以及对离群值的识别。回归分析的假设和要求经常被忽视，但不能改变这个事实：这种应予以制止的行为是极其错误的，由此所得“结果”是十分可疑的（参见 Weinzimmer 等人，1994）。下文进行一次没有识别出离群值的后续分析，并将分析结果和原始分析进行比较。SPSS 过程命令 REGRESSION 有若干选项及指令，在之后关于“回归”的章节中将详细讲述。

示例

在一次临床抽检（ $N = 711$ 名妇女）中，测量每个人的腰围和臀围。在进行分析之前，因果模型将臀围数据确定为自变量（记作“臀围”），腰围确定为因变量；对于相反的计算方向（自变量：腰围→因变量：臀围），测算出另一个有细微差别的回归函数。要预先说明的是，数据集中的 ID 号等于行号（预先设定 $ID = \$CASENUM$ ）。

预设定

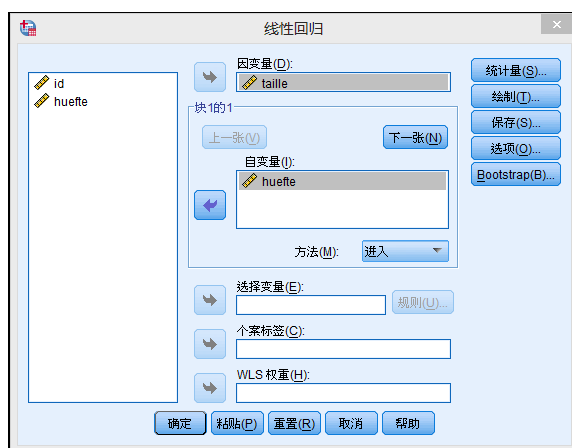
在 SPSS 程序主界面选择以下菜单项：编辑 → 选项 → “查看器”选项卡。

请确定“显示输出元素”下的选项“在日志中显示命令”已经选中。

界面操作

在 SPSS 程序主界面选择以下菜单项：分析 → 回归 → 线性...

把变量“taille（腰围）”拖入窗口“因变量”。把变量“huefle（臀围）”拖入窗口“自变量”。在“方法”下拉菜单中选择“进入”。



子窗口“统计量”：在“回归系数”一项下选定“估计值”和“置信区间”。此外，选定

“模型的拟合优度和描述性统计量”。在“残差”一项下调用“杜宾—瓦森”和带有超出3个标准差之外离群值的“个案诊断”。单击“继续”按钮。

子窗口“绘制”。针对“标准化残差”调用“直方图”和“正态分布图”。在图形方面，试着调用几个双变量散点图。如果能够统一操作，即始终在Y轴上截取标准化残差（ZRESID），并始终在X轴截取自变量或者预测的y值（ZPRED），那么随着时间的推移，解释工作就容易多了（X轴上的自变量现在只能通过语句控制才能给出说明，见下文）。通过在SPSS程序主界面选择菜单项“图形”→“旧对话框”[旧版本SPSS上没有]→“散点/点状”→“简单”分别调用散点图虽然不太方便，但是更加可靠，此外，还能利用GRAPH命令的所有功能。

在“预测值”一项下选定方差“标准化”，在“距离”一项下选定“马氏距离”、“Cook距离”和“杠杆值”，在“残差”一项下同样选定“标准化”。单击“继续”按钮。

子窗口“选项”：针对“缺失值”，选择选项“整行删除”。所选择的方法“进入”是一种逐步法，即使只有一个步骤也是如此。可以给出F值的显著性（概率）或者F值本身的数值，作为“逐步法的标准”。不要对预设进行改动。选定选项“在等式中包含常量”。单击“继续”按钮。

单击“确定”按钮开始计算。

在查看回归分析结果前，调用双变量散点图。

在SPSS程序主界面选择以下菜单项：图形→旧对话框[旧版本SPSS上没有]→散点/点状→简单...

将标准化的残差（ZRE_1，参见数据集）或者因变量（腰围）放置在Y轴上，将自变量（臀围）或者预测的y值（ZPR_1，参见数据集）放置在X轴上。

在确定图形时不仅要注意变量与轴正确的对应关系，还要注意始终将所期望的回归模型的正确残差包括在内。随着每实施一次计算过程或者存储过程，存储的变量都会增加1。最后计算得出（也就是当前的）模型的变量具有最高的末位数字（例如，ZPR_2, ZPR_3, ……）。

子窗口“标题”：为图形配上你选择的标题和脚注。

子窗口“选项”：确保为缺失值设置了和回归分析时相同的选项，在这个例子中就是“整行删除”。

如果不这样操作，数据就会有漏洞，计算和图形就会基于不同的数据，很可能会产生巨大的错误。

分别单击“确定”按钮，输出想要的图形。

调用利用语句的回归分析和图形式残差分析

查看输出语句，将其作为设定的协议并检查，是否已给定了所有需要的数据。语句必须符合下面给定的语法，只是命令行的顺序可能有所不同。但是各个选项或者整个命令行不得缺失。

在查看输出结果之前，要时常检查语法。

```
REGRESSION
  /MISSING LISTWISE
```

```

/DESCRIPTIVES MEAN STDDEV CORR SIG N
/STATISTICS COEFF OUTS CI R ANOVA
/CRITERIA=PIN (.05) POUT (.10)
/NOORIGIN
/DEPENDENT taille
/METHOD=ENTER huefte
/RESIDUALS ID (ID) DURBIN HIST (ZRESID) NORM (ZRESID)
/CASEWISE OUTLIERS (3) PLOT (ZRESID)
/SCATTERPLOT= (taille, huefte)
/SCATTERPLOT= (*ZRESID, huefte)
/SCATTERPLOT= (*ZRESID , *ZPRED)
/SAVE ZPRED COOK LEVER ZRESID MAHAL.

```

在此之后，对计算得出的模型的残差进行图形式残差分析。

Chatterjee 和 Price (1995²) 建议，为了检验模型的残差，至少要调用以下三个散点图： x * y 、 x * $y_{\text{残差}}$ 和 $y_{\text{估计值}}$ * $y_{\text{残差}}$ 。下列 GRAPH 语句分别调用三个散点图：第一个散点图含有自变量和因变量的值，第二个散点图反映自变量（“臀围”）的值和因变量“腰围”的残差；第三个散点图含有因变量（“腰围”）的估计值和残差。

```

GRAPH
/SCATTERPLOT (BIVAR) = huefte WITH taille
/MISSING=LISTWISE
/TITLE= "Linearitätstest"
/FOOTNOTE "Datenbasis: Beobachtungen".

GRAPH
/SCATTERPLOT (BIVAR) = huefte WITH ZRE_1
/MISSING=LISTWISE
/TITLE= "Test auf Varianzungleichheit und Ausreißer (1) "
/FOOTNOTE "Datenbasis: UV-Beobachtungen, AV-Residuen".

GRAPH
/SCATTERPLOT (BIVAR) = ZPR_1 WITH ZRE_1
/MISSING=LISTWISE
/TITLE= "Test auf Varianzungleichheit und Ausreißer (2) "
/FOOTNOTE "Datenbasis: AV-Schätzer, AV-Residuen".

```

关于异方差性的检验，在文献资料中也有几个针对推断性统计证明的方法，然而这些方法本身就要求不同的前提条件，随之而来也有不同的优缺点（参见 Cohen 等人著作中的总览表，2003³，130-133）。例如，根据 Darlington (1990) 提出的方法，将通过 /SAVE SRESID 命令（上面列举的程序还会在下文中予以补充）对保存的学生化残差求平方，然后通过 RANK VARIABLES 命令将其标准化，保存在所创建的变量（例如）NORMSRQ 中。

```

compute SRE1QUAD = SRE_1* SRE_1.
exe.

rank
variables= SRE1QUAD (a)
/normal

```

```

/print=yes
/ties=mean
/fraction=BLOM
/rank into NORMSRQ.

```

随后，利用 UNIANOVA 计算 SRE1QUAD 值到原始模型预测变量（包括其平方值）的回归（参考：/DESIGN=）。如果有几个预测变量，则根据 DESIGN 给定预测变量所有可能的交互作用。

```

UNIANOVA
normsrq WITH taille
/METHOD = SSTYPE(3)
/INTERCEPT = INCLUDE
/CRITERIA = ALPHA(.05)
/DESIGN = taille taille*taille

```

当在“调整模型”一行的 SPSS 输出结果中体现出显著性时，则可能就存在异方差性问题。之所以用“可能”这个词，是因为其他原因也会导致显著性。在关于异方差性的图形中，给出了这个检验的输出结果。图表和图形没有必要完全一致，优先进行哪个检验（以及为什么）必须由使用者自己决定。

回归分析语法解释

首先介绍回归分析的语法。然后解释利用 GRAPH 的图形来进行残差分析的语句。在接下来的一节，将阐述在回归分析结果之前出现的图形式残差分析结果。

REGRESSION 调用用于计算回归分析的 SPSS 过程命令。

根据 /MISSING 子命令，说明要如何处理可能含有的缺失值。LISTWISE 将所有含缺失值的行整行删除，LISTWISE 这个选项可以添加 INCLUDE 作为补充。一定要注意，在描述性分析和推断性统计分析中，对于缺失值的处理必须一致。

用/DESCRIPTIVES 子命令调用描述性统计量（MEAN，均值；STDDEV，标准差）、皮尔逊相关系数、其单侧显著性和用于计算相关（CORR，SIG，N）的个案数量。只有针对给定的变量和有效个案，才能测算描述性参数。在使用 LISTWISE 命令时，只有全部变量都具有有效数值的个案时才能进入描述性统计量的计算。

用/STATISTICS 子命令调用回归方程和自变量的统计量：R（R，R²、所显示估计值的标准误差）、ANOVA（方差分析，包括回归或残差的平方和、均方和、F 值和 F 值的显著性）、COEFF（非标准化回归系数 B、系数的标准误差、标准化回归系数（ β ）、T 和 T 的单侧显著性）、OUTS（还没有纳入方程的变量的统计量，即 βT ，T 的双侧显著性以及最小允差），CI（非标准化回归系数的 95% 置信区间）。STATISTICS 必须在 DEPENDENT 和 METHOD 之前给定。

在 /CRITERIA 子命令下，通过 PIN 决定变量根据哪些参数可以进入模型。当一个变量的统计量小于进入值时，则该变量进入模型。给定的进入值（PIN 或者 FIN）越大，则（即使是不显著的）变量就越有可能进入模型。给定的 CRITERIA 子命令必须位于 DEPENDENT 和

METHOD 子命令前面。

子命令 /NOORIGIN 使常量进入模型。/NOORIGIN 必须在 DEPENDENT 和 METHOD 之前给定。

通过子命令 /DEPENDENT 给定回归模型的因变量，例如，在这里就是变量 TAILLE。只能给定一个 DEPENDENT 子命令。必须至少有一个 METHOD 子命令跟在 DEPENDENT 子命令之后。

通过子命令 /METHOD 给定选择变量的方法，在本例中就是 ENTER。由于模型是预设定的，因此采用了直接法。根据这种方法给定预测变量，例如，在本例中是变量 “HUEFTE”。必须至少给定一个预测变量。

通过子命令 /RESIDUALS 调用第一个残差诊断。通过 ID (ID) 命令，主要是针对之后的个案诊断确定一个 ID 变量。DURBIN 选项调用杜宾—瓦森统计量。HIST 在本例中调用一个直方图，其中包括由 SPSS 创建的变量 ZRESID 的正态分布曲线图；ZRESID 包含了模型的标准化残差。通过 NORM 选项，为变量 ZRESID 调用一个 P-P 图。

利用 /SCATTERPLOT 子命令调用三个双变量散点图，由系统设定的关键词在这里被加上前缀*，以使其区别于用户设定的变量。注意：用另一个名称将这些变量存储到数据集中（参考 GRAPH 命令）。与界面选择相反，通过 /SCATTERPLOT 子命令的语句也可以给定自变量。

通过 /CASEWISE 子命令针对残差调用了范围广泛的个案诊断。通过 OUTLIERS (3)，将输出结果调整为带有 3 个标准差的标准化残差的个案。通过 PLOT (ZRESID, 预设的)，针对标准化残差调用了一个个案图形。可惜无法给定多个 /CASEWISE 命令行。

通过 /SAVE 子命令保存上面介绍的预测值、杠杆值、距离（Cook 距离、马氏距离）和残差。

图形式残差分析的语句解释

利用 GRAPH 命令创建同样的图形，即在 REGRESSION 过程中用 /SCATTERPLOT 子命令调用的图形。在此要注意两点不同：利用 GRAPHGRAPH 命令还可以给定标题和脚注；另外，GRAPH 命令在整个作用范围都可以发挥作用（尤其是模板）。在 GRAPH 命令中，首先给定 X 轴的变量，再用 /SCATTERPLOT 子命令给定 Y 轴的变量。在 GRAPH 和 /SCATTERPLOT 命令中，对于由系统计算出的相同变量需要不同的访问操作。例如，GRAPH 命令访问关于 “ZRE_1” 的标准化残差，而用 /SCATTERPLOT 命令则需要给定 “*ZRESID”。

为了调查两个变量 HUEFTE（臀围）和 TAILLE（腰围）之间的关联，利用 GRAPH 命令调用三个简单的散点图，以便通过图形式前提条件来检验计算线性回归的基本条件。

第一个图检验了在两个变量 HUEFTE（臀围）和 TAILLE（腰围）之间是否存在关联，以及这种关联是否能够通过一条直线反映出来。第一个图是根据观察数据制作的。

在本章的最后，将介绍一个 IGRAPH 实例，表示带有已绘出回归直线的一个图形。但是，回归直线在某些情况下可能产生误导。例如，当分布呈曲线形或者有离群值的时候，也会绘出回归直线；因此，在评估分布情况时应该既要使用不含有可能误导解释帮助信息的图形，又要

使用含有多个可以相互比较的描述形式的一个图，例如，SPSS 过程命令 CURVEFIT 所提供的描述形式。

第二个图检验自变量 HUEFTE（臀围）和因变量 TAILLE（腰围）标准化残差之间的关联是否具有正交性、异方差性和离群值。

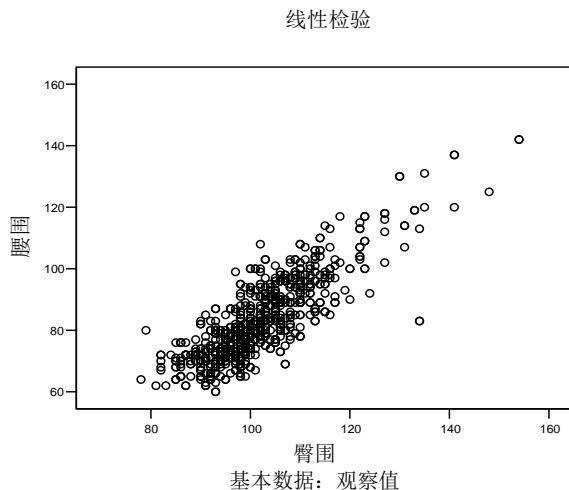
第三个图同样检验是否具有正交性、异方差性和离群值，但是根据的是估计（预测）的 TAILLE（腰围）值和因变量 TAILLE（腰围）的标准化残差。

2.1.3 输出结果和解释

图形式前提条件检验

为了检验一个模型的残差，调用了三个散点图： $x * y$ 、 $x * y_{\text{残差}}$ 和 $y_{\text{估计值}} * y_{\text{残差}}$ 。第一个散点图含有自变量和因变量的值。第二个散点图反映自变量（“臀围”）的值和因变量“腰围”的残差。第三个散点图含有因变量（“腰围”）的估计值和残差。GRAPH 和 /SCATTERPLOT 图形应得出同样的结果，因为它们是基于一个模型的相同的数据。如果 GRAPH 和 /SCATTERPLOT 图形有所不同，那么原因可能是，在 GRAPH 中设定了另一种处理缺失值的方法，或者 SCATTERPLOT 没有访问所期望模型的残差。下面只绘出和讨论 GRAPH 图形；这些讨论同样适用于 SCATTERPLOT 图形。

1. 图形：检验线性



解释：两个变量“臀围”和“腰围”之间的关联绝不是通常的云状，而是呈清晰的线性。线性的基本假设可以认为是真的。皮尔逊相关系数适合用于计算这个关联。另外，很明显这条线的走向不是与 X 轴平行的（因此斜率不等于零）。由此可以得出结论，“臀围”明显地影响变量“腰围”。在这个散点图中，数值对（135，80）（HUEFTE 臀围，约 135（观察值）与 TAILLE 腰围，约 80（观察值））的点显示为离群值。

离群值在回归分析中可能表现为两个完全不同的形式，并相应地对回归直线的估计产生两

个相反的影响。

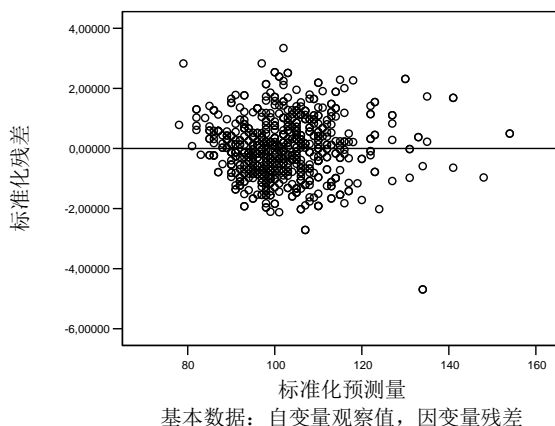
- 离群值可能与实际的线性分布呈 90° 夹角，因此或许会导致使对这种线性分布的估计部分完全出错（如本例所示）。在极端情况下，尽管呈现出线性，但仍无法估计出可用的回归方程。删除离群值可以优化对线性关联的估计。
- 离群值可能随机呈线性排列，使人觉得存在线性分布，而其他数据可能完全是离散的或者是呈点云状分布的。因此，线性是由个别离群值，而不是由大部分数据构成的。这样估计的结果就是，个别线性排列的离群值让人误以为存在线性，或者掩盖了错误的关联。在极端情况下，尽管不存在线性，但仍估计出一个线性方程。删除离群值就可以确定不存在线性，从而避免得出看似可信的线性方程。

如果有两个方差，但只有极少的离群值，如 1000 个数值中有 4~5 个离群值，就完全足以使对实际分布的估计产生偏误（当然离群值与其他数据的比例越高，偏差就越明显）。由此，回归系数、其标准误差、 R^2 和所得出结论的有效性就受到了影响。另外，在按顺序检验和删除离群值时完全有可能出现这种情况：刚开始还没有识别到真实的分布状况（图形式，至少在简单线性回归中），并且删除离群值后先是显现为非线性，而在删除更多离群值后才显现为线性。也有可能出现相反的情况。

2. 图形：检验异方差性和离群值（1）

在下面图形中，可以从这个现象识别到正交性：数值均匀地分布在零线的上方和下方。在图中，如果有这个现象就可以识别到异方差性：数值的分布呈现从左到右打开的剪刀状或者漏斗状，同时随着 X 轴值（自变量值，比如臀围）的增加，因变量（腰围）（标准化）残差的散布范围越来越大。因此，如果变量“腰围”的散布（变异性）随着变量“臀围”数值的增大而同样增大，则表明存在误差的方差不一致性（异方差性，不恒定方差）问题。

对方差不一致性和残差的检验(1)



解释：即使变量“臀围”的数值逐渐增大，变量“腰围”残差的散布（变异性）也应该保持恒定。如果变量“腰围”残差的散布（变异性）随着变量“臀围”数值的增大而增加（尤其

是其误差)，则就存在方差不一致性（异方差性）。在零线上方和下方数值的比例绝不会看起来明显不平衡，从图形上来看，也没有任何不具有正交性的迹象。数值也随机散布，也就是说，绝不会有规律地散布在零线周围，例如，不会是曲线形。

异方差性（方差不一致性）不容易识别。如果往右看，无法识别出数据散布呈剪刀状逐渐增大；但是如果往左看，数值散布呈漏斗状收敛就很明显了。这个图形表明可能有异方差性问题。在这个散点图中，一个数值对（臀围，约 135（观察值）/腰围，约-4.5（残差））所对应的点从图上显示为离群值。

组间效应检验

因变量: Rank of SRE1QUAD

来源	三型平方和	df	平方均值	F	显著性
调整模型	1234906.420a	2	617453.21	15.223	0.000
常量项	1176938.529	1	1176938.5	29.017	0.000
腰围	704498.932	1	704498.93	17.369	0.000
腰围*腰围	844291.996	1	844292.00	20.816	0.000
错误	28716729.580	708	40560.353		
总和	120060932.000	711			
调整的总变化	29951636.000	710			

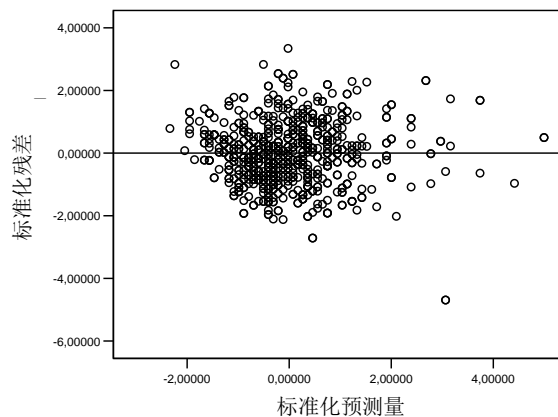
a. $R^2=0.041$ （经调整 $R^2=0.039$ ）

根据 Darlington（1990）著作的检验结果与图形结果一致。“调整模型”一行达到显著性（ $p=0.000$ ）。而且推断性统计检验也表明有异方差性问题。

3. 图形：检验异方差性和离群值（2）

第三个散点图同样检验是否具有正交性、异方差性和离群值，但是根据的是估计（预测）的估计值和各自因变量（TAILLE 腰围）的标准化残差。

对方差不一致性和残差的检验(2)



基本数据：因变量估计值，因变量残差

解释：数值呈均衡和随机的散布，也就是说，不是有规律地散布在零线周围。确定存在正交性，而且这个图形表明可能有异方差性问题。根据这些数据，最后两个图得出了同样的结论

（分别后续添加了 Y 轴上的零线）。所有散点图都应有助于识别异方差性，就好像残差的散布随着预测值或者观察值而增大。在所有散点图中，数据的散布基本都是一样的。在这里同样无法直接看出散布是规律地逐渐增大。但是如果向左看，则无法完全排除异方差性，无法明确保证具有同方差性。这些散点图表明至少有一个离群值，删除这个离群值将会得到更好的模型拟合优度（参见第 2.1.4 节）。

在这个图形式（和推断性统计式）残差分析之后，可以根据一个拟合良好的线性模型评估并相应地解释后续相关分析的结果，而不考虑可能的异方差性。对于是否具有同方差性，必要时需要更加精确地检验。

回归分析的结果

回归

在这个标题后面是所调用的线性回归的输出结果，在本例中就是一个模型，其中根据臀围数值预测出腰围数值。

描述性统计量

	均值	标准差	N
臀围	102.23	10.367	711
腰围	83.80	13.145	711

表“描述性统计量”针对有效个案（N=711）展示了均值、标准差和 N 作为模型中所有变量的描述性统计量（在这里是臀围和腰围）。

相关性

		臀围	腰围
皮尔逊相关	臀围	1.000	0.831
	腰围	0.831	1.000
显著性（1 位数）	臀围		0.000
	腰围	0.000	
N	臀围	711	711
	腰围	711	711

从表“相关性”可以看出，在两个变量之间存在很大的关联（0.831， $p=0.000$ ），因此可以计算线性回归。相关系数的绝对值越高（极差：-1~1），这种关联就越大。因此，不显著的相关系数就说明，在调查的两个变量之间存在非线性的关联或者甚至不存在关联。如果事先实施的图形式线性检验已经得出结论，两个变量之间的关联是非线性的或者曲线的，则皮尔逊相关系数不适合计算这种关联。此外，假设检验使用皮尔逊相关系数的前提条件是数据呈正态分布。

进入/剔除的变量^b

模型	进入的变量	剔除的变量	方法
1.	臀围 ^a		进入

- a. 进入了所有想要的变量。
- b. 因变量：腰围。

表“进入/剔除的变量”展示了在每个步骤进入的、或者根据方法不同重新剔除的变量的统计量。这个表格的内容取决于选择了哪种方法，做了何种设定（主要是 FIN/FOUT、TOLERANCE 等）。如果是逐步法，则“模型”一列也可以解读为“步骤”。“方法”一列反映了预设定的方法（在本例中为“进入”）；根据方法不同，在这里还有其他说明，例如，“逐步选择”和针对进入和删除设定的 F 值。由于本例采用的直接法在第一个步骤后就结束了，因此这个表格只有一行，也就是第一个（在这里也就是最后一个）模型 1。从“进入的变量”一列可以提取进入的变量，在本例中就是“臀围”。从“剔除的变量”一列可以提取剔除的变量，在本例中没有给定变量。例如，如果这个方法不剔除变量或者所有进入的变量都符合预设定的标准，则“剔除的变量”一列也是空白的。从表下面的说明可以看出，已经进入了所有想要的变量（在本例中就是“臀围”），“腰围”是模型中的因变量。

模型总结^b

模型	R	R ²	调整 R ²	估计值的标准误差	杜宾-瓦森统计量
1	0.831 ^a	0.691	0.691	7.310	1.783

a. 预测变量：（常量），臀围。

b. 因变量：腰围。

从表“模型总结”中，可以针对每个模型或者步骤获取由因变量和自变量之间关联组成的各模型最重要的特征量。在表下面，给定了模型的预测变量（又称自变量）和因变量。这个表格的内容取决于预设定了哪种方法，对于直接法通常只输出一行。对于模型 1（臀围，腰围）输出 R 、调整 R^2 、估计值的标准误差和杜宾—瓦森统计量。 R 表示因变量的观察值和模型的预测值之间线性相关的程度（极差：0~1）。如果模型中只有一个预测变量，则 R 等于表“相关”中的皮尔逊相关系数。对此的解释每次都是相同的： R 越高，模型和因变量之间的关联就越大。 $R = 0.831$ 表示模型和因变量之间存在很大的关联。 R^2 （极差：0~1）是决定系数，基于 R 值的平方。 R^2 越高，因变量中的模型对方差就解释得越好。 $R^2=0.691$ 表示，因变量中超过 2/3 的变异可以被模型解释。这个模型适合解释臀围对腰围的影响。估计值的标准误差同样是衡量模型拟合优度的一个量度。与表“描述性统计量”中的因变量（在本例中是腰围）的标准差相比较，无须了解自变量就可以得出 13.145 的标准差。但是，通过关于自变量（在本例中是臀围）可以得出小得多的估计误差，即 7.310。杜宾-瓦森统计量检验残差是否自相关，也就是说，在数值序列内部的一个残差与其前面一个残差的独立性。根据 1.783 的测定值，不能直接推断出残差的独立性具有显著性。根据杜宾-瓦森表（例如，接近 $T=750$ ， $K=2$ ， $\alpha=0.05$ ）可以确定，测定值在下方无关带之外（ $L_U = 1.87736$ ， $L_O = 1.88270$ ），因此表明正自相关具有显著性。据此，不保证残差具有独立性，无法排除自相关。但是可以通过 LAG 函数测定在数据序列中标准化残差直接（时间，空间）偏移的程度，并保存在变量 VERSETZ 中。如果这个变量与标准化残差相关，则进一步说明了如何理解自相关的显著性。

```
compute VERSETZ =lag (ZRE_1, 1) .
exe.
```

```
variable labels
ZRE_1
```

```
"Standardisierte Residuen"  
VERSETZ " 'Versetzungs'maß".  
exe.
```

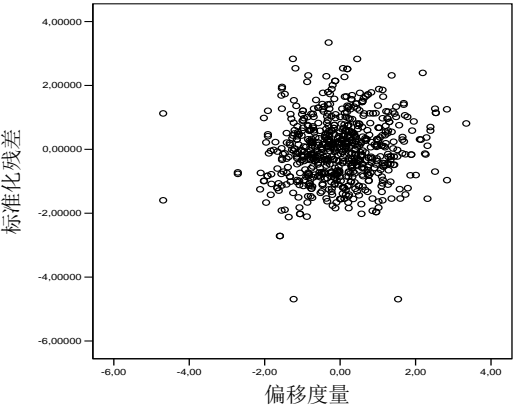
相关性

		标准化残差	偏移量度
标准化残差	皮尔逊相关	1	0.101**
	显著性 (2 位数)		0.008
	N	711	681
偏移量度	皮尔逊相关	0.101**	1
	显著性 (2 位数)	0.008	
	N	681	710

** 相关性在 0.01 的水平上 (2 位数) 是显著的。

探索性使用的皮尔逊相关表明，尽管这个关联是显著的，但从数值上来看是临界 (0.101) 的，这可能是由于样本量 (N=681) 所致。如果将标准化残差和测定的依存性偏移程度绘入散点图，则从图形上也可以看出不存在线性关联，即残差不依存于 (时间上的、空间上的) 前项数值，因此测定的杜宾-瓦森检验显著性主要是由于样本量造成的。

```
GRAPH  
  /SCATTERPLOT (BIVAR)  
  =VERSETZ WITH ZRE_1  
  /MISSING=LISTWISE .
```



散点图显示没有线性分布。据此，不能认为具有自相关，即残差 (在时间上、空间上) 依赖于前项数值。杜宾-瓦森检验的显著性是由样本量造成的。

ANOVA^b

模型		平方和	df	平方均值	F	显著性
1	回归	84804.128	1	84804.128	1587.015	0.000 ^a
	残差	37886.305	709	53.436		
	总和	122690.433	710			

a. 预测变量: (常量), 臀围。

b. 因变量: 腰围。

表“ANOVA”含有方差分析的结果，而这个方差分析的根据是，所调查模型的两个变异源“回归”（解释方差）和“残差”（未解释方差）之间的比例，相应地反映了平方和与平方均值（总和除以 df）、自由度（df）、F 值和测定的显著性。在表格下方，给定了模型的预测变量（自变量）和因变量。实施的方差分析检验了零假设——在因变量“臀围”和自变量“腰围”之间不存在线性关联。如果取得的显著性值（如表格所示）低于预设定的 α （如 0.05），则可以认为，自变量很好地解释了因变量的变异；但是，如果取得的显著性值超过预设定的 α ，则无法解释因变量的变异。根据测定的显著性，可以拒绝零假设。也就是说，在因变量“臀围”和自变量“腰围”之间存在线性关联。F 值等于变异源“回归”的平方均值除以变异源“残差”的平方均值。回归平方和的分量越大或者残差平方和的分量（观察值距离在回归直线估计值中的分量）越小，则模型对因变量变异的解释就越好。解释分量（84804.1）大于未解释分量（37886.3）。在这个背景下，也可以将 F 统计量解释为对假设的检验，即检验 R^2 是否等于零。所调查的模型适合解释因变量的变异。“回归”和“残差”一行含有关于由模型解释的和未解释的变异的详细信息。“总值”一行含有两个变异源的平方和总和与自由度总和。

回归系数^a

模型		非标准化系数		标准化系数	T	显著性	B 的 95%置信区间	
		B	标准误差	β			下限	上限
1	（常量）	-23.969	2.719		-8.815	0.000	-29.308	-18.631
	臀围	1.054	0.026	0.831	39.837	0.000	1.002	1.106

a. 因变量：腰围。

表“回归系数”针对估计模型的预测变量（在本例中是臀围）分布输出了非标准化回归系数 B（1.054）和相应的标准误差（0.026），以及置信区间（1.002 或 1.106）。此外，还给出了标准化回归系数 β （0.831）、T 值和相应的显著性。非标准化回归系数可能与标准化回归系数有很大区别，并且完全错误地表现出各自预测变量的效应（参见本例中的数值）。

根据非标准化回归系数的线性方程是：

$$\text{腰围} = -23.969 (\text{常量}) + 1.054 * \text{臀围}$$

根据标准化回归系数的线性方程是：

$$\text{腰围} = -23.969 (\text{常量}) + 0.831 * \text{臀围}$$

通常，建议将标准化回归系数用于比较在一个样本/总体内的定量变量，或者用于没有共同单位的定量变量，对于后者应考虑到，其测定取决于所选择的样本，并且根据模型拟合优度不同只能有所保留地将这种测定结果普遍化。非标准化回归系数建议用于比较样本/总体之间的定量变量，或者用于具有自然/共同单位的定量变量。根据这两种回归系数的优点和缺点，Pedhazur（1982²，247-251）建议给定两种量度。如果在分析之前将数据 z 标准化，则将 β 值等价于 B 值。

对于回归系数的解释，重要的是显著性、T 值、回归系数值和回归系数的正负号。一个变

量只有当其显著性（例如）低于 α 0.05 时，才可以用于模型。通常，只解释显著的预测变量。显著性应解释为每次检验的系数与数值零有显著区别。此外，由于输出的置信区间分别与零（说明相应的参数无效果）相距很远，这样就可以推断出，测定的参数 B 值是具有显著性的。

标准化回归系数首先对系数的解释提供了说明；回归系数越大（最大值：1），相应预测变量的影响就越大（下文将指出与解释回归系数有关联的特别之处）。标准化回归系数 >0 表明，因变量数值随着预测变量值增大而增大（线性正相关）；标准化回归系数 <0 表明，因变量数值随着预测变量值增大而减小（线性负相关）；回归系数 $=0$ 说明，相应的预测变量值完全没有影响。

根据 T 值可以看出模型中各自变量的相对意义。经验法则是， T 值应明显大于 2。带有正号（负号）的系数增大（减小）了因变量的数值。回归系数的数值与 1 相差越大，其对因变量的影响就越大。如果系数为 1，则因变量数值精确地等于自变量数值。因此在本例中，标准化正系数（0.831）应理解为线性正关联：随着臀围数值的增大，腰围也随之增大。

换言之，如果在模型中变量“臀围”增大了一个量度单位，则变量“腰围”的数值也随之增大（乘以 0.831 个量度单位）。如果在只有一个预测变量的回归模型中，回归系数等于 0，则因变量的所有数值都等于常量。

在本例中，测定的回归系数在下限（1.002）和上限（1.106）范围之内，由此可以推断，这个区间有 95% 的概率能够含有参数 B 值 1.054。SPSS 只对非标准化回归系数（ B ），而不对标准化回归系数（ β ）输出置信区间。

由于前面的分析对自变量的模型或者解释提供了比较好的数值，因此变量“臀围”也可以单独用于预测。另外，还可以通过检验（例如逐步法）是否有其他变量有助于模型做出更好的预测。

个案诊断^a

个案编号	ID	标准化残差	腰围	非标准化预测值	非标准化残差
21	21.00	-4.691	83	117.29	-34.291
392	392.00	-4.691	83	117.29	-34.291
574	574.00	3.344	108	83.56	24.443

a.因变量：腰围。

在表“个案诊断”中输出了具有标准化残差的个案，并且这些残差的标准差超过了预设的最大值 3。针对本例，显示了个案编号（活动数据集一行）、ID 以及因变量的其他数值（在本例中：腰围），尤其是非标准化预测值和非标准化残差。

“个案编号”通常不给出某个个案的 ID，而是只给出活动数据集所在行的编号。在本例中，由于 `compute ID=$CASENUM` 等于 ID 的行号，因此，在“个案编号”栏和 ID 栏显示的是同样的数值。

非标准化残差来自于因变量（在本例中：腰围）真实数值和相应非标准化预测值之间的差值。例如，在活动数据集第 21 行的个案中，腰围是 83，非标准化预测值为 117.29。相应的非标准化残差就是 $83 - 117.29 = -34.291$ 。标准化残差是残差除以其样本标准差得到的，这里平

均值为 0，标准差为 1。删除上表（个案诊断表）中的这三个个案就可以得到更好的模型拟合优度。

残差统计量^a

	最小值	最大值	平均值	标准差	N
非标准化预测值	58.26	138.37	83.80	10.929	711
标准化预测值	-2.337	4.993	0.000	1.000	711
预测值的标准误差	0.274	1.397	0.364	0.134	711
调整的预测值	58.20	138.24	83.80	10.929	711
非标准化残差	-34.291	24.443	0.000	7.305	711
标准化残差	-4.691	3.344	0.000	0.999	711
学生化残差	-4.726	3.346	0.000	1.001	711
删除残差	-34.800	24.477	0.001	7.331	711
学生化删除残差	-4.798	3.370	0.000	1.003	711
马氏距离	0.001	24.933	0.999	2.203	711
Cook 距离	0.000	0.166	0.002	0.009	711
居中杠杆值	0.000	0.035	0.001	0.003	711

a.因变量：腰围。

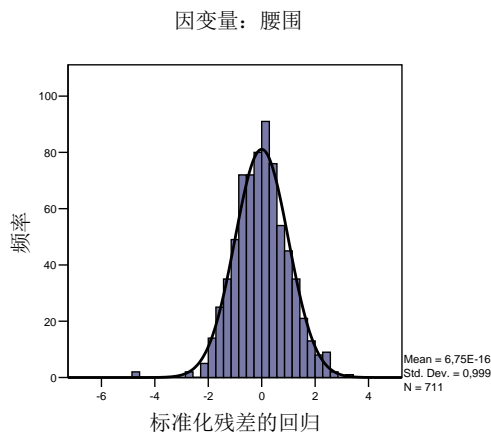
在表格“残差统计量”中，针对预测值（非标准化、标准化、预测误差、调整）、残差（差异）（非标准化、标准化、学生、删除、学生化删除）、杠杆值和高杠杆值（马氏距离、杠杆值、Cook 距离）输出了描述性统计量（平均值、标准残差、最小值、最大值、N）。标准化残差和预测值的均值为 0，标准差为 1。数值明显很大的偏差（参见最大值、最小值）表明模型适用性没有达到最佳。对其他量度，例如，杠杆值和高杠杆值的解释从细节上而言有些不一样。关于输出量度及其解释的细节可以参见模型适用性评估一节以及 REGRESSION 语句。

从“残差统计量”表无法得出残差的分布是否符合正态分布。可以从调用的图中获取这方面的信息。

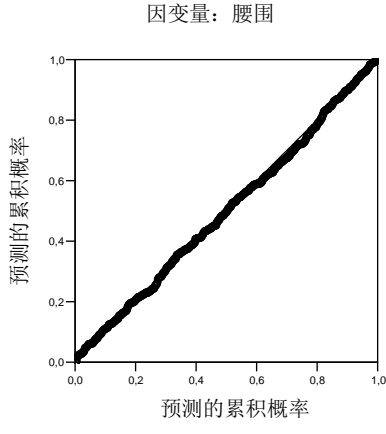
图

在输出结果中，标题“图”后面是利用 REGRESSION 调用的图。直方图和 P-P 图可以对因变量标准化残差的图形式前提条件进行检验。

直方图

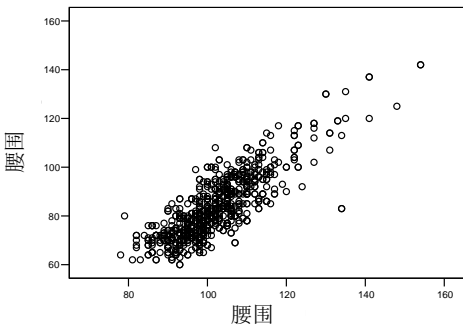


标准化残差的 P-P 图

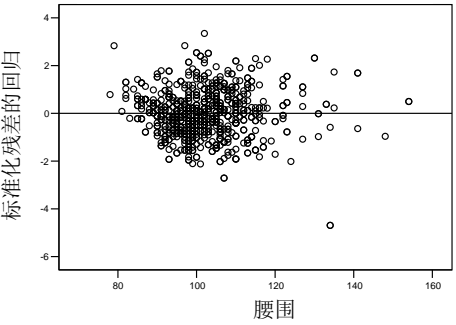


直方图表明因变量的标准化残差遵循图中所绘的正态分布；从左边可以看到一个离群值。根据直方图，可以认定标准化残差呈正态分布。P-P 图表明，因变量的标准化残差位于图中所绘的标准分布（正态分布）的直线上；相反，根据 P-P 图无法发现离群值。而且从 P-P 图上看，标准化残差也呈现正态分布。标准化残差呈正态分布时，可以视为线性回归的前提条件已经得到满足。

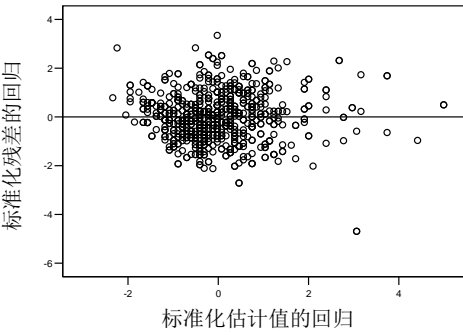
散点图
因变量：腰围



散点图
因变量：腰围



散点图
因变量：腰围



三个散点图是通过子命令 /SCATTERPLOT= 调用的，图中显示了用于识别离群值和高杠杆值 ($x * y$, $x * y_{\text{残差}}$ 和 $y_{\text{估计值}} * y_{\text{残差}}$) 的双变量分布图形，对此在前面关于残差分析的一节已介绍过了（后来又在 Y 轴添加了零线）。

模型适用性的评估

显著性或者（调整的）高 R^2 值无法保证模型可靠地与数据拟合，主要是由于这两个量都取决于 N （参见 Chatterjee & Price, 1995², 9ff, 29-32）；而且图形经常达到其可用性的极限（尤其是对于多元回归模型而言）。估计的回归系数以及其他参数对离群值和高杠杆值的反应十分敏感。在测算回归方程时，很可能有数量不多但是具有高杠杆的数据决定了分布不显著的数据量的影响。

除了前面介绍的图形式方法（如残差图）之外，为了评估观察值对多变量（回归）模型的影响，还制定了一些不同的统计量度，下面根据前面一节所计算出的双变量模型的数据对此予以解释。观察值可以描述为离群值（残差）、杠杆量度、差异量度和影响量度，对于评估多元回归的模型适用性也是非常有用的。即使事后对这些量度做正确的解释，也不能取代合理地处理数据以及进行合理性检查（参见 Rasch 等人著作，1996，571）。

杠杆作用（leverage）指的是一个个案对于估计值确定过程的影响。因此，杠杆值测定的是 X 轴上的一个点对回归拟合的影响。如果一个个案（离群值）从空间上距离分布图形其余部分（不论朝着什么方向）的中心很远，则这个个案就有很大的杠杆作用。如果一个个案远离其他数值的（线性）分布图形，则其就有很大的差异，尤其是预测 Y 值和观察 Y 值之间的残差（距离）。影响是杠杆作用和差异的后果。如果杠杆作用和差异都是比较小的尺度，则影响就很小。如果杠杆作用和差异都很大，则对个案的影响就很大。影响量度测量的方法是使一个点远离估计过程。这些数值对评估多变量模型的模型适用性非常有用，例如，多元线性回归模型。

杠杆值

杠杆作用（leverage）指的是一个个案对于估计值确定过程的影响。因此，杠杆值测定的是 X 轴上的一个点对回归拟合的影响。如果一个个案（离群值）从空间上距离分布图形其余部分（不论朝着什么方向）的中心（均值，矩心）很远，则这个个案就有很大的杠杆作用。因此，杠杆值可能位于估计直线的末端，由此施加很大的杠杆作用。杠杆作用的参数主要是马氏距离和杠杆值。马氏距离表明了，某个变量与解释变量其他个案的平均值有多大差异。如果一个个案有很大的马氏距离，则可以认为，当有一个或者几个预测变量时，这个个案具有很高的数值，因此可能对测定的回归方程有很大的影响。可以直接从中推导出两个统计量。杠杆值 = 马氏距离 / $(N-1)$ 。马氏距离 = 杠杆值 * $(N-1)$ 。杠杆值，或者说马氏距离的数值在 $0 \sim (N-1) / N$ 之间波动。这个值越大，杠杆作用越大。应调查杠杆值较高的个案（图上的点）的影响，尤其是当这些点与分布图形的其余部分有明显区别时。

检验杠杆值：

```
sort cases by LEV_1 (D) .
exe.

temp.
list variables= ID LEV_1.
```

差异值

如果一个个案（离群值）远离其他数值的（线性）分布图形，则它就有很大的差异，尤其是预测 Y 值和观察 Y 值之间的差异（距离）。差异统计量主要包括残差（参见 Cohen 等人著作，2003³，398ff.）。通过某个观察值与其相应估计值的误差，残差说明了模型适用性。对于残差图而言，残差指的是一个点与图中所绘函数直线的距离。Chatterjee 和 Price（1995²，9）建议始终要对标准化残差进行检验。根据 Cohen 等人著作（2003³，402）中的术语，标准化残差是内部学生化残差的同义词。标准化残差的平均值等于零，其标准差为 1。根据通行的规范，绝对值明显超过 3 就视为离群值或者高杠杆个案；如果绝对值超过 2，则应更仔细地进行调查。Cohen 等人（2003³，401）建议针对较大的数据量使用较高的截断点，例如 3 或者 4。使用离群值只限于检验简单（双变量）回归，不适用于多元回归模型（参见 Chatterjee 和 Price 的著作，1995²，86）。

残差检验（举例）：

```
if abs (ZRE_1) >= 2 AUSREISR=2.
exe.
if abs (ZRE_1) >= 3 AUSREISR=3.
exe.

temp.
select if AUSREISR >= 2.
list variables= ID ZRE_1.
```

输出结果示例（残差）

ID	ZRE_1
21.00	-4.69092
392.00	-4.69092
574.00	3.34377

Number of cases read: 3 Number of cases listed: 3

个案 21、392 和 574 带有示例中调用的残差（值 ≥ 3 ），在接下来的分析流程中将这几个案删除。不再阐述其他量度的输出结果和评估。

影响度量

影响是杠杆作用和差异的后果。如果杠杆作用和差异的尺度比较小，则影响就很小。如果杠杆作用和差异都很大，则个案的影响就很大。影响量度测量的是从预测过程中删除一个点对预测所产生的作用。换言之，作为一种量度，高杠杆值衡量的是如果从回归函数的测定中删除了某个个案，则所有其他个案的残差会发生多大变化。通常，每次只从预测过程中删除一个点（参见 Cohen 等人著作，2003³，402）。下面，介绍影响量度 DfFit 和 Cook 距离。这两个量度基本上是起同样作用；唯一的区别是，Cook 值不会为负值。

影响量度 DfFit（“difference in fit, standardized”的缩写）指的是由于删除某个观察值而产生的预测值变化。对于标准化 DfFit 适用的规范是，将所有绝对值大于 2 的个案除以平方根 p/N 以做检验，其中 p 是方程中自变量的数量， N 是个案的数量。Cohen 等人（2003³，404）建议对于中等或者较大的数据量使用截断点 1 或 2。Chatterjee 和 Price（1995²，89）建议，针

对这个影响量度检验所有明显很高的数值。

DfFit 统计量检验（举例）：

```
sort cases by DFF_1 (D) .
exe.
list variables= ID DFF_1.
exe.
```

针对与预测模型中的均方误差成比例的数据集，Cook 距离指的是整个数据集的预测值与减少了数据集中某一个观察值后的预测值之间的均方误差。较高的 Cook 距离表明，删除相应的个案对其余个案回归函数的计算产生了很大影响。根据通行的规范，Cook 距离超过 1 就视为具有高杠杆（参见 Cohen 等人著作，2003³，404）。Chatterjee 和 Price 在著作（1995²，89）中建议，对所有明显高的 Cook 距离进行检验。

Cook 距离检验（举例）：

```
temp.
select if COO_1 >= 1.
list variables= ID COO_1.
exe.
```

这里不再阐述其他量度的输出结果和评估。在实践中，首先带有这些个案进行回归，然后不带这些个案再进行回归，通过这种方式评估离群值和高杠杆值的意义。然后对新近开发出来的模型进行残差分析。应不仅完全从形式角度，也要从内容角度对离群值进行研究。显著数值经常是测量误差的迹象，但是也可能具有值得观注的内容。下一节进行的分析过程就从回归分析中删除了 ID 21、392 和 574。针对这个模型也进行示例中所介绍的残差分析；不能排除的可能性是，其他离群值可以被识别、检验并且在下一次分析过程中被排除。由于对残差进行了分析，因此回归分析是一种迭代分析法。

2.1.4 过程 2：删除离群值的效应——选出的输出结果

在开始计算过程之前，已经将在 2.1.3 节中识别出的离群值（ $N=3$ ）从图形分析和推断性统计分析中删除了；除了这三个个案之外，其余结果基于完全一致的数据（ $N=708$ ）和语句。

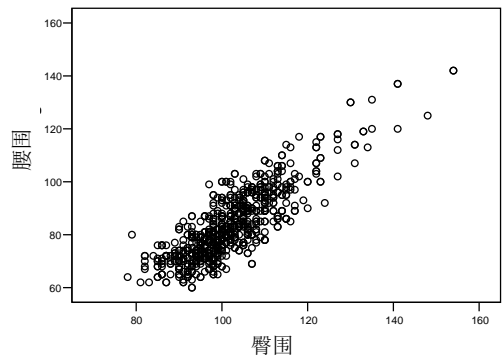
下面的讲解只限于从总样本（ $N = 711$ ）中删除三个个案而给最重要的特征量带来的变化。在 ANOVA 中，基于平方和的 F 统计量无法相互比较（因此也就没有绘出），因为它们不是基于同样数量的个案。

在分析实践中，还需利用没有离群值的新模型进行 2.1.3 节所述的（图形式）残差分析（由于篇幅所限，此处不再赘述）。就这点而言，回归分析是一种迭代过程。

图

图形法前提条件检验

第二次计算过程



与第一次进行回归分析相反，在这个散点图中再也无法发现离群值。

回归

相关性

		臀围	腰围
皮尔逊相关	臀围	1.000	0.845
	腰围	0.845	1.000
显著性（1 位数）	臀围		0.000
	腰围	0.000	
N	臀围	708	708
	腰围	708	708

皮尔逊相关系数从 0.831 改为 0.845。

模型总结^b

模型	R	R ²	调整 R ²	估值的标准误差	杜宾-瓦森统计量
1	0.845 ^a	0.714	0.714	7.028	1.769

a. 预测变量：（常量），臀围。

b. 因变量：腰围。

R 从 0.831 变为 0.845，R²从 0.691 变为 0.714。

系数^a

模型		非标准化系数		标准化系数	T	显著性	B 的 95%置信区间	
		B	标准误差	β			下限	上限
1	（常量）	-26.916	2.647		-10.168	0.000	-32.113	-21.719
	臀围	1.084	0.026	0.845	42.026	0.000	1.033	1.134

a. 因变量：腰围。

非标准化系数从 1.054 变为 1.084。标准化系数从 0.831 变为 0.845。

在某些应用领域（社会科学、生物计量学），这些临界变化可能是无关紧要的；但是对于含有重要信息的计量经济学建模而言，这些边际变化具有深远的后果。

在这里不再复述其他统计量（如离群值、残差）和图。

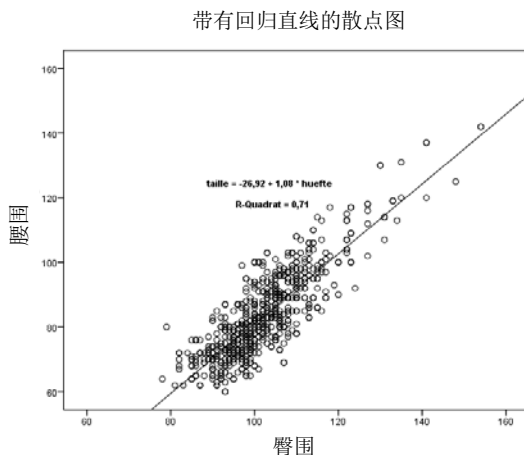
2.1.5 说明：绘制回归直线（IGRAPH）的图形

通过 SPSS 过程命令 IGRAPH 也可以调用绘制了回归直线的散点图。但是，绘出的直线回归函数可能有误导作用。例如，当分布呈曲线形或者具有离群值时，也会绘出直线（参见 Chatterjee 和 Price 的著作，1995²，10）。

语句：

```
IGRAPH
/TITLE="Streudiagramm mit Regressionsgerade"
/X1 = VAR (huefte) TYPE = SCALE
/Y = VAR (taille) TYPE = SCALE
/COORDINATE = VERTICAL
/FITLINE METHOD = REGRESSION LINEAR LINE = TOTAL MEFFECT SPIKE=OFF
/SCATTER COINCIDENT = NONE.
```

图形



记录的回归方程不是基于标准化回归系数，而是基于非标准化回归系数。

2.2 非线性简单回归

非线性线性模型的线性化

不是所有函数都能满足简单线性回归的前提条件。本节将尽可能简单地从线性回归过渡到非线性回归，并帮助解答下列问题：

- 如果两个变量之间的关联是非线性的怎么办？
- 如果双变量线性回归模型的残差不是随机分布，而是非线性分布怎么办？
- 如果在一个双变量线性回归模型中存在异方差性怎么办？

在调查两个变量的相关时，可能有不同迹象表明存在非线性关联（参见 Chatterjee 和 Price 的著作，32-33）。

1. 一种概念上的原因可能是：理论模型从一开始就假定不存在线性关联，而是存在非线性关联。

如果一个模型的参数是非线性的，则存在固有非线性模型。固有非线性模型不适合用 OLS 回归进行分析（参见 Pedhazur 的著作，1982²，404-405）。

2. 根据经验的事实结果可能是：对线性模型残差的检验得出了一个结论：完全不存在线性关系，而是存在非线性关系，因此，为了可以使用常见的线性方法，必须对模型进行转换。

如果一个模型的参数是线性的，但是变量是非线性的，则就存在固有线性模型。通过适当地转换可以将固有线性模型线性化，并利用 OLS 回归进行分析。

3. 一种概率理论上的原因可能是：对于因变量而言，均值和方差相互关联。此时不满足方差一致性，在这里出现了异方差性，导致产生不准确的估值。

异方差性违背了线性模型中方差的一致性。一般通过简单地转换就可以消除异方差性。

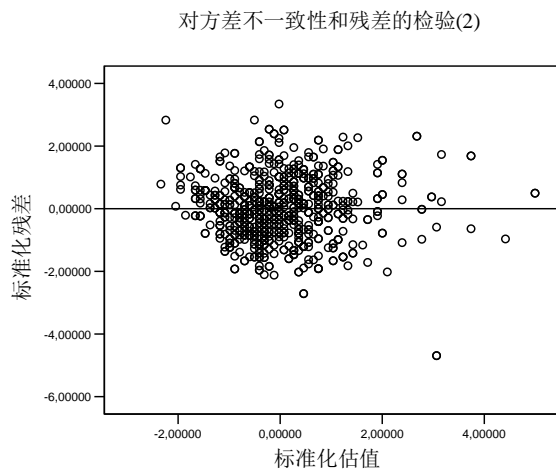
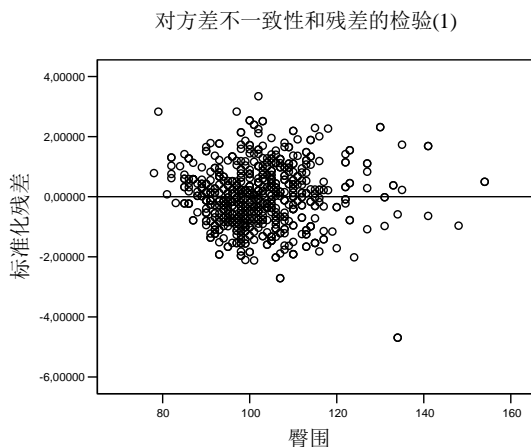
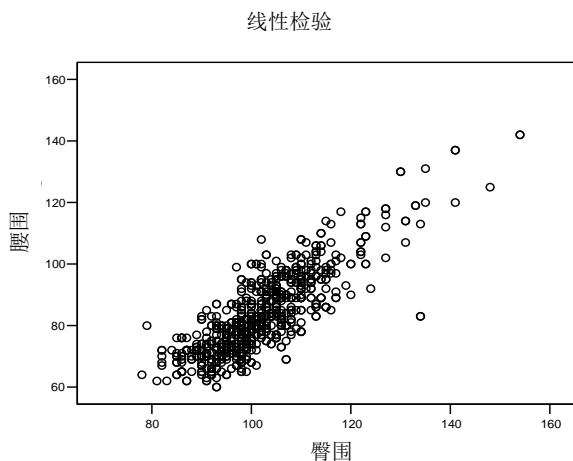
这些问题并不总是很容易识别和解决。但是在这里强烈建议，不要试图利用线性回归模型或者随机出现的、随意选择的函数来“近似地”描述或者解释没有线性化的固有线性函数或者固有非线性函数。

原则上，非线性回归的方法与线性回归的方法没有区别。这两者的共同前提条件是，在开始对其分析之前，模型基于（双变量）的函数必须是已知的。因此，（非）线性回归并不是一种数据结构化的方法，而是描述和检验模型是否与给定的数据结构拟合。在实践中，对此的简化表达如下：

- 如果利用线性回归调查线性模型，则观察值应线性分布，而模型的（标准化）残差应在零线周围大致呈随机分布（参见第 2.2.1 节）。
- 如果利用线性回归调查非线性（例如，对数）模型，则模型的残差呈非线性分布（参见第 2.2.2 节）。
- 但是，如果对非线性（例如，指数）模型的变量进行适当的转换，然后用线性回归进行调查，则线性化模型的残差在转换后应在零线周围呈随机分布（参见第 2.2.3 节，固有线性函数的方差）。
- 固有非线性模型能且只能应用非线性回归进行调查（参见第 2.2.4 节。关于非参数回归请参见 Cohen 等人著作，2003³，252-253）。

2.2.1 利用线性回归对线性函数进行分析

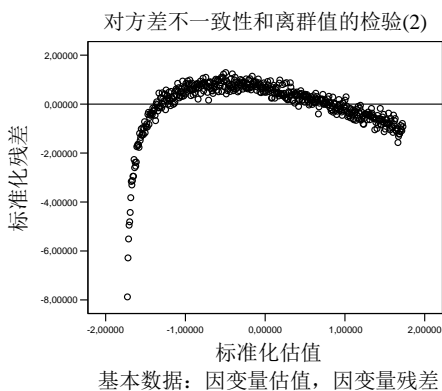
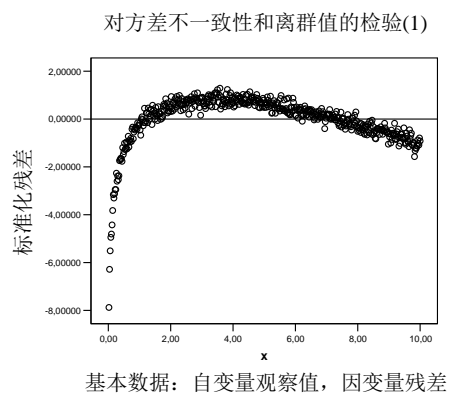
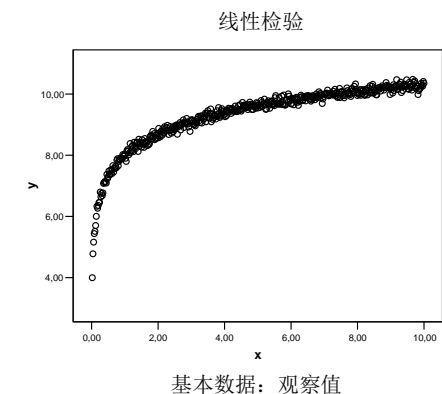
如果利用线性回归调查线性模型，则残差的曲线走向应符合两个变量之间关系的模型。换言之，在线性回归中，数值的分布必须是线性的，残差的方差必须相同。因此，在散点图中，观察值应是线性的，模型的残差应随机地分布在零线周围。下图来自对简单回归的概述。



满足了线性、正交性和方差一致性的前提条件。回归方程是适合和精确的，在删除了离群值之后更是如此。

2.2.2 利用线性回归分析调查非线性函数

如果利用线性回归调查非线性（如对数）函数，则模型的残差呈非线性分布，在这个个案中就是对数分布。



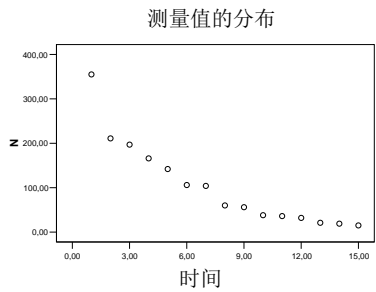
观察到这里没有给定线性、没有给定正交性、残差不等于零且散布很大。所调查分布情况的特性表明，这些数据不适合利用线性回归进行分析。估计的线性回归方程不适合非线性分布，很不可靠（后来在 Y 轴上添加了零线）。

2.2.3 将非线性函数线性化，并利用线性回归进行调查

但是，如果对非线性（如指数）模型的变量进行适当的转换，然后用线性回归进行调查，则线性化模型的残差在转换后应在零线周围呈随机分布。这些图的数据摘录自 Chatterjee 和 Price 的著作（1995²，38-39）。

因变量的对数化是一种解决异方差性问题的常用技术，但是也会导致出现散布减小和不对称现象（Chatterjee 和 Price 的著作，1995²，53）。

在 X 光照射下的细菌死亡率

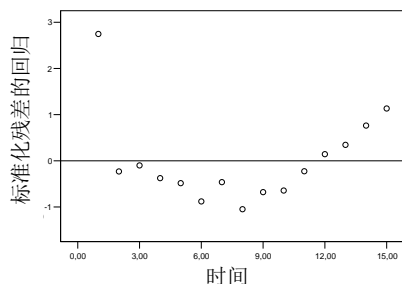


* (a) 原始值的图形 *

```
GRAPH
/SCATTERPLOT (BIVAR)=zeit WITH n
/MISSING=LISTWISE
/TITLE= 'Bakteriensterblichkeit
bei Röntgenbestrahlung'
'Verteilung der Messwerte'
/FOOTNOTE= 'Quelle: Chatterjee & Price,
1995'.
```

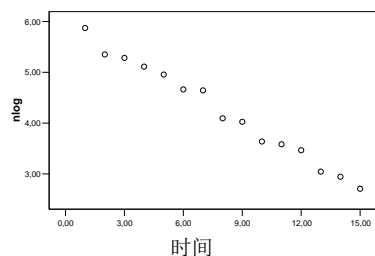
散点图

因变量: N



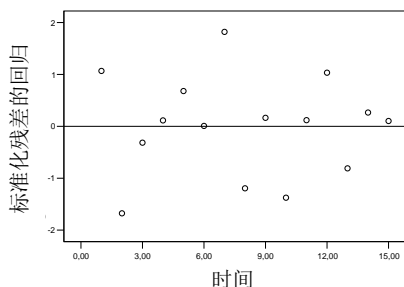
在 X 光照射下的细菌死亡率

已转换数值的分布



散点图

因变量: nlog



* (b) 对基于未转换数值的非线性分布的线性分析*.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN (.05) POUT (.10)
  /NOORIGIN
  /DEPENDENT n
  /METHOD=ENTER zeit
  /SCATTERPLOT= (*ZRESID,zeit)
  /RESIDUALS NORM (ZRESID) .
```

* (c) 因变量的转换*.

```
compute nlog=ln(n) .
exe.
```

* (d) 已转换数值的图形*.

```
GRAPH
  /SCATTERPLOT (BIVAR)=zeit WITH nlog
  /MISSING=LISTWISE
  /TITLE= 'Bakteriensterblichkeit
  bei Röntgenbestrahlung'
  'Verteilung der transformierten
  Werte'
  /FOOTNOTE= 'Quelle: Chatterjee &
  Price, 1995'.
```

* (d) 线性化分布的回归分析*.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN (.05) POUT (.10)
  /NOORIGIN
  /DEPENDENT nlog
  /METHOD=ENTER zeit
  /SCATTERPLOT= (*ZRESID,zeit)
  /RESIDUALS NORM (ZRESID) .
```

在转换后, 线性化模型的残差线性、正态和随机地分布在零线周围。在将非线性分布线性化后, 满足了线性回归的前提条件。估计的回归方程是稳定和精确的。将固有线性函数线性化后, 就可以使用普通最小二乘估计法。

在这里可能出现的唯一问题就是对回归方程的解释，经常会出现一个双重问题。尽管转换后的因变量和未转换的自变量之间的关联是直接和线性的。可是，有时运用和理解比较麻烦的数学转换并不是一件容易的事情。但是，这些结果是通过在数值相关中进行转换得到的，如果要将其向回推导到所研究关联的经验相关，可能就变得十分困难。换言之，在回归方程中不再有原先提到的变量，而是只有经过转换的变量。因此，未做改动的变量“细菌死亡率”和“时间”之间的关联可能比经过对数化处理的“细菌死亡率”和“时间”之间的关联更容易理解。例如，如果经过对数化处理的细菌死亡率随着时间下降，究竟意味着什么？因此，在对不同拟合良好的回归方程进行比较时，不仅是尽量最佳的参数（例如， R^2 等），而且相对简单的解释也是具有决定性意义的，而这种解释是理想地基于对原始变量尽量简单的转换。只有用户自己，而且只有根据期望水平、应用关联和对后果的估计才能在回归方程的精确性和实用性之间做出抉择（参见关于应用目的影响和回归方程评估标准的下一节）。

2.2.4 利用非线性回归分析非线性函数：非线性回归

可以利用非线性回归调查非线性模型。非线性回归是基于迭代的估计算法，在这里根据模型不同，可以选择 Levenberg-Marquardt 算法（NLR，没有伴随条件的模型）或者序列二次优化算法（CNLR，带有伴随条件的模型）。

非线性回归的方法不能与 WLS 法（Weighted Least Squares；加权最小二乘法）混淆，后者同样可以用于估计非线性回归模型。WLS 法是通过将加权的残差平方和最小化进行参数估计的。OLS 法是通过将没有加权的残差平方和最小化进行参数估计的。在使用 WLS 法时，根据与扰动项方差倒数的比例选择加权。根据 Chatterjee 和 Price 的著作（1995²，53），对于经过转换的变量 y/x 和 $1/x$ 而言，使用 WLS 法和 OLS 法没有什么区别。

初学者在这里可能提出的唯一问题是：“好的，那么我的（双变量）分布是根据哪个函数呢？”受过数学教育的研究人员，例如，数学家、物理学家或者统计学家经常可以仅仅观察一条分布曲线就得出一个比较恰当的函数方程。例如，函数是升序还是降序的，是指数的还是对数的等。为了清楚地阐明在非线性回归时的处理方法，我们在概述时首先使用一个已知线性函数的回归方程。

由于预测过程是迭代的，因此语句控制在透明度和效率方面可能超过鼠标控制（Schendera，2005）。为了使计算出的回归方程具有更加直观的可理解性和可检验性，下面就对各种编程方式做出解释。

个案 1 函数已知（本例：线性函数）

根据关于简单线性回归的一章，回归方程是已知的。根据非标准化回归系数的线性方程是：“腰围” = $-23.969 + 1.054 * \text{“臀围”}$ 。经过四舍五入后，给定这些（已知）参数，作为 MODEL PROGRAM 一项下的初始值。了解回归方程所涉及的调查对象，对于确定参数和初始值是非常有帮助的。如果初始值选择的太差，那么尽管模型函数正确，还是有可能出现一种很坏的情况，即固有算法做出输出没有收敛、内容毫无意义的估计（参见后面的余弦示例），或者不是提供全面最佳的，而是只提供局部最佳的答案。第 2.2.6 节将进一步解释 NLR 语法。所需的方程摘录自第 2.2.4 节的总览表“CURVEFIT 函数：名称、要求、方程和残差”。

语句：

```

MODEL PROGRAM b0 = -23.0 b1 = 1.0 .
COMPUTE NPRED = b0 + b1*huefte.
NLR taille
/PRED=NPRED
/SAVE RESID PRED
/CRITERIA ITER 100 SSCONVERGENCE 1E-8 PCON 1E-8 .

```

解释说明

通过 **MODEL PROGRAM** 命令，给定回归方程参数（效应）的粗略估计值和各自的初始值，在这个个案中就是已知回归方程中经过四舍五入的效应和非标准化回归系数。通过 **COMPUTE** 命令给定回归方程本身，在这个个案中就是已知的线性回归方程；预测变量的数值保存在变量 **NPRED** 中。**NLR** 调用非线性回归的计算。根据 **NLR** 命令，从模型中给定因变量，根据 **/PRED** 命令再次给定应包含预测值的因变量（**NPRED**）。通过 **SAVE** 命令，保存残差（**RESID**）和预测值（**PRED**）。根据 **CRITERIA** 子命令可以给定迭代算法的设定和标准，尤其是迭代次数和中断标准。细节可参见关于 **NLR** 和 **CNLR** 语句的后续解释。

非线性回归分析

参数估计值

参数	预测变量	标准误差	B 的 95%置信区间	
			下限	上限
b0	-23.969	2.719	-29.308	-18.631
b1	1.054	0.026	1.002	1.106

借助于 **NLR** 语句，用非线性方法对一个线性函数建模。测定了与用线性方法进行分析时一样的回归方程（参见“参数估计值”表），也就是带有常量（**b0**）-23.97 和斜率参数（**a** 或 **b1**）1.05。只有利用 **COMPUTE** 命令给定的回归方程，才能对其函数关系进行解释。由于这里是线性关系，因此也可以根据线性回归对 **b0** 和 **b1** 进行解释。但是，通常只有在极个别情况下才能像线性回归一样对非线性回归的参数进行解释。下一节解释其他表。

个案 2 函数未知（本例：非线性函数）**利用 CURVEFIT 的曲线估计**

尽可能最佳的模型拟合优度，是计算线性（非线性）回归的基本前提之一。如果身边没有能力很强的统计学家，就可以使用 **SPSS** 的 **CURVEFIT** 函数。**SPSS** 过程命令 **Prozedur** 提供了一种简便的方法，可以十分清晰地解释两个变量关联的函数。

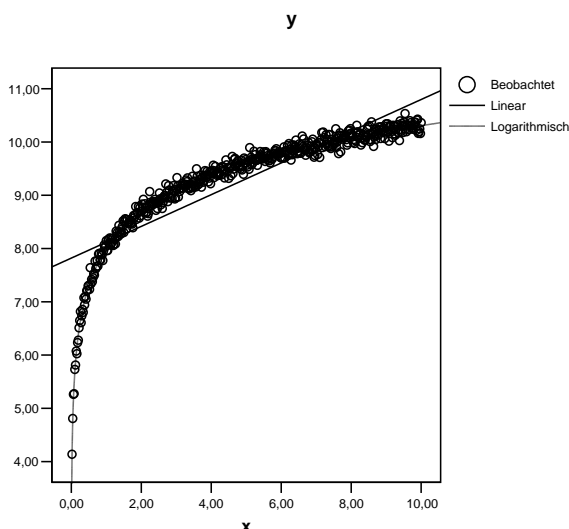
利用 CURVEFIT 的模型拟合

CURVEFIT 不仅检验可能存在的线性关联，而且还检验另外 10 个关联模型（主要是指数、指数分布、逆、立方、对数、二次、**S (S)**、增长和复合）。从根据经验存在的成对测量值（“观察”）的排列中，**CURVEFIT** 截取测算出的直线函数（如果一次性绘制大量函数，就有

可能看不到全貌)。此外, CURVEFIT 针对每个函数都测算出统计参数, 例如, 显著性、 R^2 等。因此不仅可以通过肉眼观察直线, 还可以根据统计参数对不同的曲线模型进行比较。此外, 通过比较还可以相对简单地判定哪个函数可能比直线模型更好地反映出所调查的两个变量之间的关联。这里需要注意的是, 应用目的也同时确定了在创建回归方程时需要优化的标准。

在 SPSS 程序主界面选择以下菜单项:
分析 → 回归... → 曲线估计 → 定义...

将 Y 确定为因变量, X 确定为自变量。从下面提供的模型中选择想要的曲线函数。选中“在等式中包含常量”和“模型的图”复选框。单击“确定”按钮, 调用曲线估计。



```
TSET NEWVAR=NONE .
CURVEFIT
/VARIABLES= y WITH x
/CONSTANT
/MODEL=
  LINEAR LOGARITHMIC INVERSE
  QUADRATIC CUBIC COMPOUND
  POWER S GROWTH EXPONENTIAL
  LGSTIC
/UPPERBOUND=11
/PLOT FIT .
```

为了让图形明了一些, 在这里只绘出了线性函数和对数函数。其他函数将在后面的章节通过其参数予以介绍。

解释说明

CURVEFIT 调用曲线估计的方法。CURVEFIT 标准化地输出了一个曲线估计图和回归统计量的综合表, 其中主要包括曲线函数或曲线方法、 R^2 、自由度、 F 值、显著性水平、上限 (Upper bound)、常量 (b_0) 和回归系数 (b_1 、 b_2 、 b_3)。置信区间预设定为 95%。CURVEFIT 标准化地整行删除缺失值。

在 VARIABLES 后面是曲线函数的两个变量的名称, 数据都应针对这些曲线函数进行调整。提到的第一个变量 (在这里是 X) 作为自变量被列入建模; 只能给定一个自变量。然后提到的另一个变量 (在这里是 Y) 构成了因变量; 可以给定几个因变量。如果 X 成为因变量, Y 成为自变量, 则结果会得出其他函数 (对此参见简单线性回归的说明), 但 R^2 还是同一个。只能给定一个 VARIABLES 命令。/CONSTANT 命令决定了回归方程是否应包含一个常量 (或者是否不包含: NOCONSTANT)。根据 /MODEL= 命令, 一次性可以给出最多 11 个不同的回归模型 (也可以通过选项: ALL) 实现。由于 CURVEFIT 命令针对每个因变量和模型曲线都自动创建 4 个新的变量, 因此在数据集很大时不应再将其作为必要的曲线估计而调用。下面详细介绍不同的回归模型。如果调用一个逻辑回归模型 (LGSTIC), 则必须利用 /UPPERBOUND 单独给出一个上限值, 这个值是正数并且大于所给出的所有因变量中的最大

值；对于现有数据，已经给出了数值 11。针对逻辑回归模型，在输出结果中给出这个上限值。利用 PLOT FIT（预设定）命令调用曲线估计图，用 PLOT NONE 命令删除曲线估计图的输出结果。

在 CURVEFIT 前面的 TSET NEWVAR= 命令确定了处理回归统计量的预设定。在 NONE 时检验回归统计量，而不存储预测值和残差值；相反，在 CURRENT 和 ALL 时存储预测值和残差值，此时 CURRENT 代替先前的变量，ALL 则不代替。预设定的例子不适用于时间序列数据。

输出结果

模型总结和参数估计值

因变量: y

方程	模型总结					参数估计值			
	R^2	F	自由度 1	自由度 2	显著性	常量	b1	b2	b3
线性	0.768	1650.680	1	498	0.000	7.831	0.296		
对数	0.989	46150.544	1	498	0.000	8.010	0.996		
逆	0.607	263.200	1	498	0.000	9.456	-0.206		
二次	0.646	2271.498	2	497	0.000	7.029	0.775	-0.048	
立方	0.646	2897.764	3	496	0.000	6.477	1.433	-0.212	0.011
复合	0.633	1077.833	1	498	0.000	7.806	1.035		
幂函数	0.633	25667.595	1	498	0.000	7.903	0.121		
S (S)	0.620	395.667	1	498	0.000	2.245	-0.028		
结构函数	0.633	1077.833	1	498	0.000	2.055	0.034		
指数	0.633	1077.833	1	498	0.000	7.806	0.034		
逻辑	0.643	4932.228	1	498	0.000	0.040	0.815		

自变量是 x。

解释说明

已经如上所述绘出了曲线估计图。在根据经验存在的成对测量值（散点图）的排列中，针对每个调用的函数绘出一条直线（线性的、对数的等）。如果将各条直线的比较限制在基本数据的值域内（SPSS 使绘出的直线超过作为基础的值域，这实际上是不允许的），则可以根据曲线估计图的图形确认，哪几个测算出的函数将会得出几乎相等的结果；在这里不要将“相等”与“同样好”混淆。为了更有说服力地在函数之间做出选择，可以首先查看表格中的回归统计量。

针对所调查的模型（参见总览和图例）和分别创建的曲线函数或者曲线方法，表“模型总结和参数估计值”列出了对模型的总结和估计的参数。在“方程”一项下给定了分别创建的曲线函数或者曲线方法，在“模型总结”一项下给定了 R^2 、F 值、自由度 1 和 2 以及获得的显著值 [“Sig.”]。此外，先前的 SPSS 输出结果还包括设置的上限（“Upper bound”）。在“参数估计值”一项列出了各自模型的参数：常量（b0）、b1、b2 和 b3。

如果没有达到允差标准,则在这个表格下面显示一条提示。可以通过 TSET 调整 QUA 和 CUB 的允差标准。很明显,当 R^2 为 0.989 时对数函数的表现最佳。逆函数 ($R^2=0.346$) 得出了模型拟合优度的最差 R^2 值。

选择回归方程 (模型拟合优度的标准)

根据输出的参数,选择回归方程的前提条件首先是,变量选择要适合回归方程的应用目的。根据曲线估计图可以认为,绘出的直线表现了对观察到的成对测量值的合理估计。误差散布应最小,或者在理想情况下所有的点都在函数上,这一点在对数函数的图中比线性函数的图中体现得更明显。

对于模型适用性的重要提示是, x 值函数与最大值和最小值范围拟合 (从图形上来看:直线的始端和末端与 X 轴最大值和最小值处于同一高度)。如果这些数值指向的方向和数据不相同,则表明函数没有较好地与数据拟合,不适合进行预测 (对此参见下文关于余弦函数的例子)。如果测定的曲线完全没有反映出根据经验存在的或者预期的分布,则即使是最好的参数也是毫无意义的。例如,如果数据清楚地显示出单调的上升趋势,则绝不能选择呈现下降趋势的曲线。

还有几个评估函数适用性的要素。除了前面已经提到的最高点、最低点和单调性 (上升或下降) 外,还有曲线的转折点、零点 (与 X 轴的切点) 和曲率。可能有的间断 (在曲线和 (或) 数据中) 一方面表明有取样错误,但是也可能是来自组成的数据和 (或) 函数。

下一个标准是显著性。首先观察使 F 检验达到显著性的曲线函数 (在本例中是所有模型)。另一个标准是 R^2 (还有其他标准: s^2 和 C_p ; 参见 Chatterjee 和 Price 的著作, 1995², 246 页)。从显著的模型中,只观察 R^2 值最高的曲线函数 (在本例中只有 LOG)。

回归方程的相对简单性是下一个标准。例如,更好的模型参数 (例如 R^2) 经常要面对更复杂的回归方程,但是其复杂性无法总是由较小的模型拟合优度数值的改善而抵销。例如,如果一个线性回归模型和一个三次回归模型的 R^2 只有 0.001 的差别,但是对此的代价是在二次函数的方程中增加了两个变量 (参见总览表),那么在这种情况下,可能优先选择更简单的方程,也就是线性回归方程。

这样做的优点,不仅是可以继续利用线性相关模型或者线性回归模型进行计算而没有实质性的信息丢失,而且利用线性模型对计算结果或者模型的解释,比利用二次模型的解释更加简单。但是最后应再次确认,用图形和统计方法找到的曲线函数是否真正适合描述相互关联的构件。

为了进行最后的观察,经常应进行两方面的试验。一是如果延长所找到的函数,从而超过现有数据的值域可能会导致什么样的后果;二是根据值域不同,公共函数是否无法分解为可能不相同的单个函数。

把识别到的函数应用于数据

现在,通过 CNLR 命令使利用 CURVEFIT 测定的函数与数据拟合。CNLR 命令的语句和输出结果符合第 2.2.4 节个案 1 所述的、除了附加条件 “ $b_0 \geq 0$ ” 外的操作。第 2.2.6 节将进一步对语法进行解释。所需的方程摘录自第 2.2.4 节的总览表 “CURVEFIT 函数: 名称、要求、方程和残差”。

语句:

```
MODEL PROGRAM b0=8.0 b1=1.0 .
COMPUTE NPRED = b0+ln (b1*x) .
CNLR Y
/PRED=NPRED
/SAVE RESID PRED
/BOUNDS b0 > 0
/CRITERIA STEPLIMIT 2 ISTEP 1E+20 .
```

输出结果

带有附加条件的非线性回归分析

迭代记录^b

迭代 ^a	残差平方和	参数	
		b0	b1
1.0	5.109	8.000	1.000
1.1	1.589E8	-562.085	571.090
1.2	1401894.858	-49.006	58.011
1.3	7228.312	2.302	6.703
1.4	12.283	7.433	1.573
1.5	5.097	7.946	1.060
2.0	5.097	7.946	1.060
2.1	5.112	7.836	1.178
2.2	5.095	7.935	1.072
3.0	5.095	7.935	1.072
3.1	5.095	7.913	1.096
3.2	5.095	7.933	1.075
4.0	5.095	7.933	1.075
4.1	5.095	7.929	1.079

通过数字计算衍生变量。

- a. 初始迭代的数字显示在小数点左边，次级迭代的数字显示在小数点右边。
- b. 在完成了 14 次模型评估和 4 次衍生变量评估后停止回归分析，因为前后紧接着的残差平方和之间的相对缩小最高可以达到 $SSCON = 1.00E-008$ 。

表“迭代记录”给出了估计过程的步骤。在“迭代”一列，小数点左边的数值给定了初始迭代的数字，小数点右边的数值给定了二次迭代（次级迭代）的数字。这样，在四次初始迭代后找到了答案。其他列含有每个迭代步骤的残差平方和与参数估计值。

参数估计值

参数	预测变量	标准误差	95%置信区间	
			下限	上限
b0	-7.933	169934.814	-333849.973	333865.839
b1	1.075	182600.920	-358762.070	358764.220

表“参数估计值”归纳了每个参数的预测值。在这个例子中，由于在分析之前参数就是已知的，通过 CURVEFIT 命令已经很好地对参数做了估计，并将其引入回归方程，因此取得的估计值是接近完美的。对于非线性回归的参数，没有必要和线性回归的参数一样予以解释。在本例中，b0 相当于常量，b1 相当于经过对数化处理的斜率参数。如果 95%置信区间的上限和下限与数值零（零：相关参数无影响）相差很远，则可以得出参数是显著的结论。

参数估计值的相关系数

	b0	b1
b0	1.000	-1.000
b1	-1.000	1.000

表“参数估计值的相关系数”反映了参数的内相关性。如果只有当样本足够大时（“渐近”）标准的参数估计值相关系数矩阵才表现出很高的相关性，则应检验模型是否过度参数化、模型参数是否过多以及数据状况是否不佳。在本例中，则表明试图从回归方程中删除参数 b1，因为它几乎等于 1 了。

ANOVA^a

来源	平方和	df	平方均值
回归	43863.769	2	21931.885
残差	5.095	498	0.010
未调整的总值	43868.864	500	
调整的总值	476.538	499	

因变量: y

a. $R^2 = 1 - (\text{残差平方和}) / (\text{调整的平方和}) = 0.989$ 。

表“ANOVA”含有方差分析的结果。对所调查模型的变异源“回归”（解释方差）和“残差”（未解释方差）的解释与线性回归时的类似输出结果一致。但是不输出 F 值和显著性。“未修正总值”相当于总变异性，也就是“回归”和“残差”之和。“调整总值”表现了“平均”y 值的变异性。通过用 1 减（“残差”/“调整总值”）计算“R²”（R²），R² 为 0.989，表示模型解释了大约 98.9%的因变量变异性。

与 OLS 回归相反，在非线性回归中不输出推论统计量。因此在非线性回归中，只有根据很大的样本才能测定可靠的，也就是渐近的标准误差和置信区间。

通过 CLNR 并利用非线性回归，检验了利用 CUR-VEFIT 测定的非线性函数是否与数据拟合。我们已经看到，用常量（b0）8.00 和斜率参数（a 或 b1）1.00 测定出了完全相同的回归方

程。渐近相关系数矩阵表明，参数 b1 可能是多余的，因为它几乎等于 1，因此即使将其从回归方程中删除也不会有明显的信息丢失。

CURVEFIT 函数：名称、要求、方程和残差

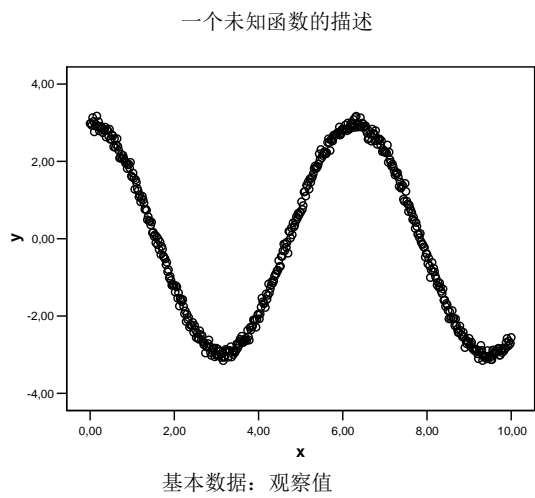
选项	方程	线性方程	对数化残差
[LIN]EAR 线性	$y = b_0 + b_1 * x$	$y = b_0 + b_1 * x$	
[LOG]ARITHMIC 对数	$y = b_0 + b_1 * \ln(x)$	$Y = b_0 + b_1 * \ln(x)$	
[INV]ERSE 逆	$y = b_0 + b_1 / x$	$y = b_0 + b_1 / x$	
[QUA]DRATIC 二次	$y = b_0 + b_1 * x + b_2 * x^2$	$y = b_0 + b_1 * x + b_2 * x^2$	
[CUB]IC 立方	$y = b_0 + b_1 * x + b_2 * x^2 + b_3 * x^3$	$y = b_0 + b_1 * x + b_2 * x^2 + b_3 * x^3$	
[COM]POUND 复合	$y = b_0 * b_1^x$	$\ln(y) = \ln(b_0) + x * \ln(b_1)$	COMPUTE NEWVAR = LN (VAR) - LN (FIT#n)
[POW]ER 幂	$y = b_0 (x^{b_1})$	$\ln(y) = \ln(b_0) + b_1 * \ln(x)$	COMPUTE NEWVAR = LN (VAR) - LN (FIT#n)
S	$y = e^{b_0 + b_1 / x}$	$\ln(y) = b_0 + b_1 / x$	COMPUTE NEWVAR = LN (VAR) - LN (FIT#n)
[GRO]WTH 增长	$y = e^{b_0 + b_1 x}$	$\ln(y) = b_0 + b_1 * x$	COMPUTE NEWVAR = LN (VAR) - LN (FIT#n)
[EXP]ONENTIAL 指数分布	$y = b_0 (e^{b_1 x})$	$\ln(y) = \ln(b_0) + b_1 * x$	COMPUTE NEWVAR = LN (VAR) - LN (FIT#n)
[LGS]TIC Logistic	$y = (1 / (u + b_0 b_1^x)) - 1$	$\ln(1/y - 1/u) = \ln(b_0) + x * \ln(b_1)$	COMPUTE NEWVAR = LN (VAR) - LN (1/FIT#n) 或者具有给定的上限： COMPUTE NEWVAR = LN (1/VAR - 1/u) - LN (1/FIT#n)

图例。Y/VAR：因变量。X：自变量或者时间值。B0：常量。Bn：回归系数。e：欧拉数：2.71828）。Ln/LN：自然对数。u：LGSTIC 时的上限。NEWVAR：对数化残差。FIT#n：由 CURVEFIT 创建的拟合变量的名称。

对于模型 COMPOUND、POWER、S、GROWTH、EXPONENTIAL 和 LGSTIC 而言，如果因变量中的数值小于或者等于 0，则无法进行对数转换。为了进一步回归诊断，可以通过 COMPUTE 命令测定对数化残差。SPSS V13 资料中的 POWER 等式是不正确的。

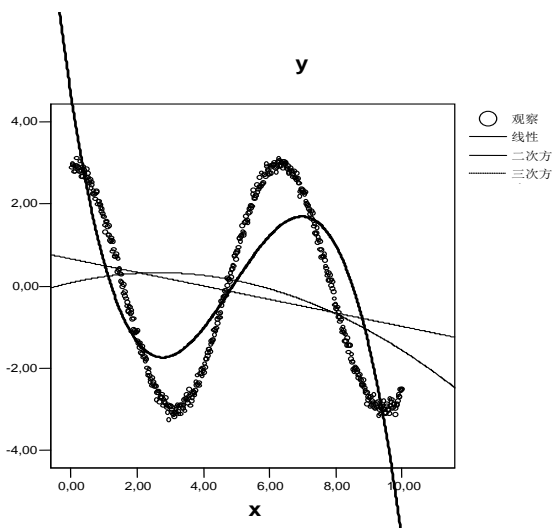
个案 3 所寻找的函数不包括在 SPSS 过程命令 CURVEFIT 中：CURVEFIT 的含义和极限

如果所寻找的函数既不是已知的，又不包括在 CURVEFIT 中，则可以请有经验的统计学家界定所寻找的函数。下面这个例子选择了一个 CURVEFIT 没有提供的函数。第一个步骤是描述数据的分布。



```
GRAPH
  /SCATTERPLOT (BIVAR) = x
WITH y
  /MISSING=LISTWISE
  /TITLE= "Beschreibung
einer unbekannten Funktion"
  /FOOTNOTE "Datenbasis:
Beobachtungen".
```

下一步是通过 CURVEFIT 尝试第一个近似值。



CURVEFIT 只能使线性、二次和三次函数与数据拟合（参见下文）。为了便于查看，三次函数的曲线超出了图的边框。

```
TSET NEWVAR=NONE .
CURVEFIT
/VARIABLES=y WITH x
/CONSTANT
/MODEL=
  LINEAR LOGARITHMIC INVERSE
  QUADRATIC CUBIC COMPOUND
  POWER S GROWTH EXPONENTIAL
  LGSTIC
/UPPERBOUND=4
/PLOT FIT .
```

CURVEFIT 只能使线性、二次和三次函数与数据拟合，因此，对于调用的所有其他函数不再输出参数。

输出结果

模型总结和参数估计值

因变量: y

方程	模型总结					参数估计值			
	R^2	F	自由度 1	自由度 2	显著性	常量	b1	b2	b3
线性	0.050	26.212	1	499	0.000	0.671	-0.167		
对数 ^a	0.000	0.000		
逆 ^b	0.000	0.000		
二次	0.064	16.914	2	498	0.000	0.108	0.171	-0.034	
立方	0.715	415.010	3	497	0.000	4.672	-5.333	1.343	-0.092
复合 ^c	0.000	0.000		
幂函数 ^a	0.000	0.000		
S (S) ^b	0.000	0.000		
结构函数 ^c	0.000	0.000		
指数 ^c	0.000	0.000		
逻辑 ^c	0.000	0.000		

自变量是 x。

- a. 自变量 (x) 含有不是正数的数值, 最大值是 0.00。无法计算对数模型或者幂模型。
- b. 自变量 (x) 含有为零的数值。无法计算逆模型或者 Holt 模型。
- c. 因变量 (y) 含有不是正数的数值, 最小值是-3.15。无法使用 Log 转换。复合的结构模型、幂模型、Holt 模型、结构模型、指数模型和对数模型不能用于这个变量。

解释说明

已经如上所述绘出了曲线估计图。根据调用的函数, 只能将线性、二次和三次的回归方程与现有数据拟合。因此, 表中只列出了 LIN、QUA 和 CUB 的参数。根据 CURVEFIT, 三次回归方程 (CUB) 的表现看上去是最好的 ($R^2=0.715$)。

如果选择了这个结果, 就犯了一个错误。仔细观察散点图就会发现, 在最右边根据经验采集的观察值的波形线重新上升, 相反三次函数的曲线继续下降。因此, 数据和三次函数的分歧越来越大。如果求助于统计学家, 当他看到波形的数据图形后可能会这么说: “哦, 就我看到的曲线而言, 可能是振幅为 3 的余弦函数。”

现在, 对函数 $y = \cos(3x)$ 进行了基于 NLR 语句的非线性回归, 然后检验方差解释的程度。第 2.2.6 节进一步解释了 NLR 语句, 其中还包括必要的方程。

语句:

```
MODEL PROGRAM b1=3 .
COMPUTE NPRED = b1*cos (x) .
NLR y
```

```
/PRED=NPRED
/SAVE RESID PRED
/CRITERIA ITER 100 SSCONVERGENCE 1E-8 PCON 1E-8 .
```

输出结果

非线性回归分析

迭代记录^b

迭代 ^a	残差平方和	参数
		b1
1.0	5.152	3.000
1.1	5.139	2.993
2.0	5.139	2.993

通过数字计算衍生变量。

- a. 初始迭代的数字显示在小数点左边，次级迭代的数字显示在小数点右边。
- b. 在完成了 3 次模型评估和 2 次衍生变量评估后，停止回归分析，因为前后紧接着的参数估计值之间的相对缩小最高可以达到 $PCON = 1.00E-008$ 。

参数估计值

参数	估计值	标准误差	95%置信区间	
			下限	上限
b1	2.993	0.006	2.981	3.005

ANOVA^a

来源	平方和	df	平方均值
回归	2349.334	1	2349.334
残差	5.139	500	0.010
未调整总值	2354.473	501	
调整总值	2340.313	500	

因变量: y

a. $R^2 = 1 - (\text{残差平方和}) / (\text{调整的平方和}) = 0.998$ 。

解释说明

R^2 （“R 方”）达到 0.998，因此明显超过三次模型函数的模型拟合优度（如方差解释）。

如果将利用 CURVEFIT 测定的三次回归方程进行基于 NLR 语句的非线性回归（所需的方程摘录自 2.2.4 节的总览表“CURVEFIT 函数：名称、要求、方程和残差”），则方差解释的程度不发生改变， R^2 基本保持不变（0.715）。其原因绝不是测定方程的方式（也就是通过 CURVEFIT，另一种方法是通过 NLR 语句；先前的 SPSS 版本中在这里可能会出现细微差别），

而是由于三次函数与数据不拟合，而且应用的 SPSS 过程命令对此也无法改变。

示例：三次函数转化为 NLR

```
MODEL PROGRAM b0=4 b1=-5 b2=1 b3=-0.1.
COMPUTE NPRED = b0+ (b1*x) + (b2*x*x) + (b3*x*x*x) .
NLR y
/PRED=NPRED
/SAVE RESID PRED
/CRITERIA ITER 100 SSCONVERGENCE 1E-8 PCON 1E-8 .
```

结果（节选）

ANOVA ^a			
来源	平方和	df	平方均值
回归	1689.386	4	422.347
残差	668.979	497	1.346
未调整总值	2358.365	501	
调整总值	2344.835	500	

因变量: y

a. $R^2 = 1 - (\text{残差平方和}) / (\text{调整的平方和}) = 0.715$ 。

CURVEFIT 的意义可能就在于，方便地使标准函数与现有分布状况拟合，并且只要在数据中隐藏了标准函数，就方便地测定数据基于的回归方程。当然必须考虑到，分布状况不包括在 CURVEFIT 中的函数。但是，CURVEFIT 自身的极限可能从图形中反映出了所估计的函数。自动分配（截尾）的轴长度有可能给定了根本不存在的曲线走向，过于稠密的点有可能掩盖了曲线函数的拐点（对此参见关于余弦函数的上例）。

同样，测定的回归方程 R^2 可能起到误导作用。即使是方差解释达到将近 70%，也不表明选择的函数正确。因此，如果函数是未知的，应请教了解所调查对象的、有经验的统计学家。

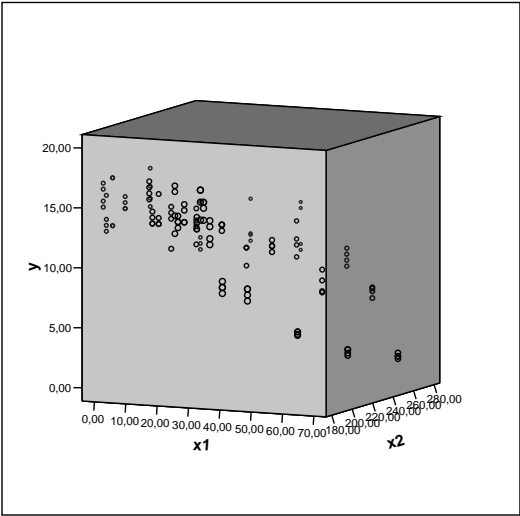
市场上有专用的曲线估计程序，可以针对带有两个或者多个预测变量的模型将上千个函数拟合。但是经验表明，在其他程序中对所输出的方程进行检验是非常必要的。原因是，对于可以评估所输出方程优度的准则（如 R^2 ），程序并不总是一并输出的。例如，在另一个程序（如 SPSS）中进行检验可能就得出这样的结果：一个由程序利用达到最佳的卡方统计量输出的方程，在 SPSS 中只能得出基本及格的 R^2 值。

2.2.5 更高的要求：带有两个预测变量的非线性回归

之前讲述的非线性模型只含有一个预测变量。但是，SPSS 可以拟合所含预测变量超过一个的模型。演示所用的数据和回归方程是根据对两个预测变量时模型能力下降现象的分析（Nelson, 1981）。预测变量是时间（单位：周）和温度（单位：℃），因变量测量了绝缘极限值的强度（单位：千瓦），模型函数是已知的。

利用三维立方体对测量值分布进行探索，表明了两个预测变量 x_1 （时间，单位：周）和 x_2

(温度, 单位: °C) 的影响。



模型公式:

回归方程是:

$$\log_y = \beta_1 - \beta_2 x_1 * \exp(-\beta_3 x_2) + e。$$

```
compute YLOG=ln (Y) .  
exe.  
  
MODEL PROGRAM b1=2.5 b2=0.000000005 b3=-0.05 .  
COMPUTE NPRED = b1- (b2*x1*exp (-b3*x2) ) .  
NLR ylog  
  /PRED=NPRED  
  /SAVE RESID PRED  
  /CRITERIA ITER 100 SSCONVERGENCE 1E-8 PCON 1E-8 .
```

非线性回归分析

迭代记录^b

迭代 ^a	残差平方和	参数		
		b1	b2	b3
1.0	48.490	2.500	5.000E-9	-0.050
1.1	6.035E16	2.590	-4.043E-8	-0.113
1.2	5.702	2.480	1.184E-8	-0.055
2.0	5.702	2.480	1.184E-8	-0.055
2.1	3.981	2.594	8.344E-9	-0.056
3.0	3.981	2.594	8.344E-9	-0.056
3.1	4.462	2.591	5.018E-9	-0.058
3.2	3.801	2.593	7.445E-9	-0.057
4.0	3.801	2.593	7.445E-9	-0.057
4.1	3.923	2.591	5.416E-9	-0.058

续表

迭代 ^a	残差平方和	参数		
		b1	b2	b3
4.2	3.799	2.593	7.258E-9	-0.057
5.0	3.799	2.593	7.258E-9	-0.057
5.1	3.799	2.592	6.860E-9	-0.057
6.0	3.799	2.592	6.860E-9	-0.057
6.1	3.800	2.591	6.111E-9	-0.057
6.2	3.798	2.592	6.687E-9	-0.057
7.0	3.798	2.592	6.687E-9	-0.057
7.1	3.798	2.592	6.339E-9	-0.057
8.0	3.798	2.592	6.339E-9	-0.057
8.1	3.798	2.591	6.013E-9	-0.057
9.0	3.798	2.591	6.013E-9	-0.057
9.1	3.798	2.591	5.704E-9	-0.058
10.0	3.798	2.591	5.704E-9	-0.058
10.1	3.798	2.591	5.619E-9	-0.058
11.0	3.798	2.591	5.619E-9	-0.058
11.1	3.798	2.591	5.618E-9	-0.058
12.0	3.798	2.591	5.618E-9	-0.058
12.1	3.798	2.591	5.618E-9	-0.058

通过数字计算衍生变量。

a. 初始迭代的数字显示在小数点左边，次级迭代的数字显示在小数点右边。

b. 在完成了 28 次模型评估和 12 次衍生变量评估后停止了回归分析，因为前后紧接着的残差平方和之间的相对缩小最高可以达到 $SSCON = 1.00E-008$ 。

表“迭代记录”给出了预测过程的步骤。这样，就在 12 次初始迭代后找到了答案。

参数估计值

参数	估计值	标准误差	95%置信区间	
			下限	上限
b1	2.591	0.019	2.553	2.629
b2	5.618E-9	0.000	-6.480E-9	1.772E-8
b3	-0.058	0.004	-0.066	-0.050

表“参数估计值”反映了每个参数的估计值。由于在本例中，参数真实值在分析之前是已知的，并且已经引入了回归方程，因此所获得的预测值与其基本相等。由于 95%置信区间的上限和下限距离零（相关参数无效应）很远，可以得出除 b3 之外的参数是显著的结论。

参数估计值的相关性			
	b1	b2	b3
b1	1.000	0.451	0.442
b2	0.451	1.000	1.000
b3	0.442	1.000	1.000

从表“参数估计值的相关性”可以看，参数 b3 可以完美地与 b2 相关，因此可以将其从回归方程中删除。

ANOVA ^a			
来源	平方和	df	平方均值
回归	716.368	3	238.789
残差	3.798	125	0.030
未调整总值	720.165	128	
调整总值	54.413	127	

因变量：YLOG

a. $R^2 = 1 - (\text{残差平方和}) / (\text{调整的平方和}) = 0.930$

带有两个预测变量的模型解释了因变量大约 93%的变异性。

2.2.6 用于非线性回归的 SPSS 过程 NLR 和 CNLR

为了计算非线性回归，SPSS 提供了 NLR 和 CNLR 两个函数，它们使用迭代算法，原则上适合估计因变量和自变量之间的任意关系。

NLR 和 CNLR 的区别

NLR 和 CNLR 的主要区别在于，CNLR 可以给定参数的附加条件（也就是 Constraints，参见 /BOUNDS），因此在估计算法（如 ISTEP）方面有所不同。NLR 的优点是其用法比 CNLR 简单，因为不需要给定附加条件。下面对同一个非线性回归的编程进行比较，一个是用 NLR，另一个是用 CNLR，两个例子都是基于同样的模型和数据。为了将其表现出来，表格中的输出结果用文本表示。

语句：

```
NLR          MODEL PROGRAM
              b0=5.0  b1=-0.20 .
              COMPUTE NPRED = b0+b1*ZEIT.
              NLR nlog
              /PRED=NPRED
              /SAVE RESID PRED
```

```

/CRITERIA ITER 100
SSCONVERGENCE
1E-8 PCON 1E-8.

```

输出结果（节选）：

非线性回归

R squared = 1 - Residual SS / Corrected SS = .98836

Parameter	Estimate	Asymptotic Std. Error	Asymptotic 95 % Confidence Interval	
			Lower	Upper
B0	5.973160267	.059778098	5.844017539	6.102302996
B1	-.218425255	.006574714	-.232629062	-.204221449

语句：

```

CNLR      MODEL PROGRAM
           b0=5.0  b1=-0.20 .
           COMPUTE NPRED = b0+b1*ZEIT.
           CNLR nlog
           /PRED NPRED
           /SAVE RESID PRED
           /BOUNDS B0 >= 0
           /CRITERIA STEPLIMIT 2
           ISTEP 1E+20.

```

输出结果（节选）：

带有附加条件的非线性回归

R squared = 1 - Residual SS / Corrected SS = .98836

Parameter	Estimate	Asymptotic Std. Error	Asymptotic 95 % Confidence Interval	
			Lower	Upper
B0	5.973160267	.059778098	5.844017539	6.102302996
B1	-.218425255	.006574714	-.232629062	-.204221449

如果在 CLR 中给定了一个无效的附加条件，那么这两个命令可能得出完全一致的结果（尽管 CNLR 具有一个附加条件，即 $B0 > 0$ ）。在上面这个例子的数据中，参数 B0 从一开始就大于零。因此，相当于在 CNLR 中给定了一个根本不需要的附加条件。CNLR 和 LNR 的结果完全一致。

对于所介绍例子的 NLR 和 CNLR 语句

本节介绍了在前面例子中使用的 NLR 和 CNLR 语句。详细情况和更多信息请查阅“SPSS

Command Syntax Reference”（SPSS 命令语句参考）一书。

MODEL PROGRAM 命令

利用必要的 MODEL PROGRAM 命令给定所有的参数及其初始数值，这些是之后给定的（非）线性回归方程（从 NLR 和 CNLR 命令开始）所需要的。这些参数包括在分析函数时需要使用的常量项、乘法系数、指数或数值。MODEL PROGRAM 命令必须位于 NLR 或 CNLR 命令之前。原则上可以给定任何名称，但是，根据总览表，在此建议遵守关于方程表述的规范，即如 B0、B1 等。如果在 MODEL PROGRAM 命令下忘记了一个参数，则不执行 NLR 或 CNLR 命令。对于参数，原则上可以给定任意数值。在给定数值 0 或 1 时应特别小心，因为这些任意数值既可能代表一个常量（1），也可能不包含信息（0）。

COMPUTE 命令

在 COMPUTE 命令下给定（非）线性回归方程，以便能够计算因变量的预测值。因变量的预测值定义了（非）线性模型。回归方程必须符合表述（非）线性方程的规范，因此除了自变量之外，必须至少还包括 MODEL PROGRAM 命令中的一个参数。对于含有预测值的变量，原则上可以给定任何名称，预设定的名称是 PRED。在上面这些例子中，使用的名称是 NPRED。根据所需的回归方程不同，不必给定 COMPUTE 命令，作为 COMPUTE 命令的代替，可以给定很多其他命令，如 IF、DO IF、END IF、LOOP、END LOOP 或者 COUNT。

CNLR / NLR 命令

根据用途不同，必须给定 NLR 或者 CNLR 命令。在 NLR 或 CNLR 命令的后面，直接给定（非）线性回归的因变量。只能给定一个来自活动数据集的数值因变量。

根据/PRED 子命令给定含有预测值的变量。该变量的名称必须与在回归方程中确定的名称一致。在这些例子中给定的名称是 NPRED，因此根据/PRED 子命令给定变量名称 NPRED，以表示这个变量含有因变量的预测值。只有利用 SAVE 子命令保存了在这里定义的变量，才能将预测值存储在工作数据集中。

SAVE 子命令

利用 SAVE 子命令可以将预测值、残差和模型衍生变量的临时变量永久地存储在工作数据集中，以便在后面的分析中检验模型的拟合优度。根据 SAVE 命令，必须至少给定下列关键词中的一个。

PRED 保存因变量的预测值。

RESID 保存残差。在括号内可以给定自变量名称。例如，如果根据 RESID 给定了“（RESX）”，则将数据集的残差保存在名称“RESX”下面。

DERIVATIVES 保存所有衍生变量。

LOSS（只用于 CNLR，并且只有给定了 /LOSS 子命令时才有）保存用户定义的损失函数（loss function）的变量。非线性回归中的损失函数是由算法最小化的函数。

关键词的顺序和编排并不重要。但是，如果在打开的数据集中已经有了所需保存变量的名称，则不进行保存。

CRITERIA 子命令

利用 CRITERIA 子命令可以设置迭代算法，尤其是迭代次数和停止标准。由于 CNLR 和 LNR 的算法不同，因此在 CRITERIA 项下通常给定不同的选项。

LNR 的选项

LNR 使用了 Levenberg-Marquardt 算法，对于这种算法可以给定几个停止标准：迭代最大次数、平方和收敛和参数收敛。只要迭代算法达到了给定的其中一个停止标准，就找到了最佳答案。

ITER n: 迭代的最大次数。对于 ITER，可以给定任意的正整数值；预设定为每个参数 100 次。

SSCONVERGENCE n: 平方和的收敛标准。对于 SSCONVERGENCE 可以给定不是负数的任意数值；预设定为 1E-8。如果前后连续的迭代无法根据这个比例使平方和发生改变，则停止使用该算法。如果给定 0，则该标准失效。

PCON n: 参数收敛的标准。对于 PCON，可以给定不是负数的任意数值；预设定为 1E-8。如果前后连续的迭代没有根据这个分量使其中一个参数值发生改变，则该 SPSS 过程命令停止。如果给定 0，则该标准失效。

CNLR 的选项

CNLR 使用序列二次优化算法，对于这种算法可以给定迭代最大次数、步距、最优允差、函数精度和无限步距。如果达到了最优允差，则找到了最佳答案。

ITER n: 迭代的最大次数。对于 ITER，可以给定任意正整数值。

STEPLIMIT n: 步距。步距 n 防止优化算法远离良好的初始预测值。对于 STEPLIMIT，可以给定任意的正整数值。预设定的数值是 2。

ISTEP n: 无限步距 n。无限步距指的是定义为无限的参数所发生变化的程度。只要参数变化的程度大于 ISTEP 标准，则停止估计。对于 ISTEP，可以给定任意的正整数值；预设定的数值是 1E+20。

BOUNDS 子命令（只 CNLR 中有）

根据 BOUNDS 子命令，可以给定 MODEL PROGRAM 命令参数的附加条件。一个附加条件就是在迭代地寻找答案时对一个或者几个参数数值的限制。根据 BOUNDS 子命令，可以给定一个或者几个（用分号分开）线性或非线性附加条件。附加条件可能复杂程度不同，从简单的单变量条件到十分复杂的多变量条件。在 BOUNDS 子命令下，只能给定 MODEL PROGRAM 命令参数的附加条件，只能使用算术运算符和关系运算符。关于非线性附加条件的说明请查阅“SPSS Command Syntax Reference”（SPSS 命令语法参考）一书。

2.2.7 非线性回归的假设

了解回归方程所涉及的调查对象，在确定函数和迭代算法的初始值时是非常有帮助的。

1. 非线性模型的线性化对比非线性回归。在利用 LNR 或者 CNL 进行非线性回归之前，

应检验非线性模型是否可以转化为线性模型，然后用 OLS 回归（SPSS 过程命令 REGRESSION）对其进行分析。

- 2. 变量。因变量和自变量必须是定量的。只有分类变量事先转换成了二元变量，才可以将其纳入非线性回归。
- 3. 函数。只有当确定的回归方程（函数）精确地吻合因变量和自变量之间的关联，也就是所有数据点都在绘出的函数图形上时，非线性回归的结果才是可靠的。一个函数应既没有超出现有的测量值域，又不应在测量值域之外被解释。
- 4. 初始值。对于迭代算法，应尽量选择所需预测模型的参数初始值，因为初始值会影响收敛。初始值应尽量与预计的模型函数最终答案一致。

根据散点图可以得到需要建模的参数的第一批近似值。根据需要建模的函数不同，最小值、最大值和各变量之间的特定差值或者比例是十分有用的。如果初始值选择的太差，那么尽管模型函数正确，还是有可能发生迭代算法根本不收敛、输出内容毫无意义的估计或者只得出局部最佳而不是整体最佳的答案。为了实现成功的收敛，必要时应给定附加条件。

- 5. 迭代次数。迭代次数应足够多，从而在达到最大次数之前找到答案。因此，如果由于达到最大迭代次数而停止迭代，那么此时测定的模型可能不是最好的模型。如果尽管迭代很多次还是没有找到答案，则应修改初始值或者回归方程。
- 6. 样本范围。只有当样本范围足够大时，非线性回归的结果才是可靠的。与 OLS 回归相反，在非线性回归中不输出推论统计量。因此，在非线性回归中，只有根据很大的样本才能测定可靠的标准误差和置信区间。
- 7. 巨大数值。巨大数值乘方的模型在某些情况下可以测定过大或者过小的数值，以便将这些数值显示出来。必要时，可以通过适当的初始值或者附加条件防止产生这种上溢或者下溢现象。

2.2.8 总览表：非线性回归的模型

下面的总览表列出了不同非线性模型的几个例子。这个表格是根据参数数量（b1—n）和自变量数量排列的。尤其是对于多值方程，表格中列出的方程通常只是几种方程表述形式中的一种。所选择的、使用固有名称的方程不应起到误导作用，让人以为除此之外还有很多“无名”的函数。

	函数的名称和说明	
参数数量	一个自变量（x）	模型公式
1	常量模型	b1（常量，a）
	通过原点的直线	b1*x
	通过原点的抛物线	b1*x**2
	双曲线	b1/x

续表

参数数量	函数的名称和说明	
	一个自变量 (x)	模型公式
2	直线	$b_1 \cdot x + b_2$
	抛物线	$b_1 \cdot x^2 + b_2$
	幂函数	$b_1 \cdot x^{b_2}$
	指数	$b_1 \cdot \exp(b_2 \cdot x)$
	超几何	$b_1 \cdot x^{b_2} \cdot (b_2 \cdot x)$
	对数	$b_1 + b_2 \cdot \ln(x)$
	余弦	$b_1 \cdot \cos(b_2 \cdot x)$
	正弦	$b_1 \cdot \sin(b_2 \cdot x)$
	正切	$b_1 \cdot \tan(b_2 \cdot x)$
3	Ellipse	$\sqrt{b_1 - b_2 \cdot (x - b_3)^2}$
	Hoerl	$b_1 \cdot (b_2 \cdot x) \cdot x^{b_3}$
	Michaelis-Menten	$b_1 \cdot x / (x + b_2)$
	渐近回归	$b_1 + b_2 \cdot \exp(b_3 \cdot x)$
	β	$b_1 \cdot (x^{b_2} \cdot (1 - x)^{b_3})$
	Cauchy	$1 / (b_1 \cdot (x + b_2)^2 + b_3)$
	密度	$(b_1 + b_2 \cdot x)^{-1} \cdot (-1 / b_3)$
	Freundlich	$b_1 \cdot x^{b_2} \cdot (b_2 \cdot x^{b_3})$
	Gauss	$b_1 \cdot (1 - b_3 \cdot \exp(-b_2 \cdot x^2))$
	Gamma	$b_1 \cdot (x / b_2)^{b_3} \cdot \exp(-x / b_2)$
	Gompertz	$b_1 \cdot \exp(b_2 \cdot \exp(b_3 \cdot x))$
	Gunary	$x / (b_1 + b_2 \cdot x + b_3 \cdot \sqrt{x})$
	Harris	$1 / (b_1 + b_2 \cdot x^{b_3})$
	Johnson-Schumacher	$b_1 \cdot \exp(-b_2 / (x + b_3))$
	Langmuir	$b_1 / (b_2 + x^{b_3})$
	Logistic	$b_1 / (1 + b_2 \cdot \exp(b_2 \cdot x))$
	Log-Modified	$(b_1 + b_3 \cdot x)^{b_2}$
	Log-Logistic 对数成长	$b_1 \cdot \ln(1 + b_2 \cdot \exp(-b_3 \cdot x))$
	Metcherlich Law of Diminishing Returns	$b_1 + b_2 \cdot \exp(-b_3 \cdot x)$
	Verhulst	$b_1 / (1 + b_3 \cdot \exp(-b_2 \cdot x))$
	收益密度	$(b_1 + b_2 \cdot x + b_3 \cdot x^2)^{-1}$
4	Morgan-Mercer-Florin	$(b_1 \cdot b_2 + b_3 \cdot x^{b_4}) / (b_2 + x^{b_4})$
	Ratio of Quadratics	$(b_1 + b_2 \cdot x + b_3 \cdot x^2) / (b_4 \cdot x^2)$
	Richards	$1 / ((b_1 + b_2 \cdot \exp(b_3 \cdot x))^{b_4})$
	Von Bertalanffy	$(b_1 \cdot (1 - b_4) - b_2 \cdot \exp(-b_3 \cdot x))^{b_4} / (1 - b_4)$
	Weibull	$b_1 \cdot \exp(b_2 \cdot x^{b_3}) + b_4$

续表

	函数的名称和说明	
参数数量	一个自变量 (x)	模型公式
5	Peal-Reed	$b1 / (1 + b2 * \exp (- (b3 * x + b4 * x^{**2} + b5 * x^{**3})))$
	Ratio of Cubics	$(b1 + b2 * x + b3 * x^{**2} + b4 * x^{**3}) / (b5 * x^{**3})$
	两个自变量 (x1, x2)	
2	颠倒双曲线	$x1 / (b1 + b2 * X2) ;$ 或者: $x2 / (b1 + b2 * X1)$
	超几何	$b1 * x1^{**} (b2 * x2) ;$ 或者: $b1 * x2^{**} (b2 * x1)$
3	Hoerl	$b1 * (b1^{**} x1) * x2^{**} b3 ;$ 或者: 等。
	Freundlich	$b1 * x1^{**} (b2 * x2^{**} b3)$
	Gamma	$b1 * ((x1 / b2)^{**} b3) * \exp (x2 / b2)$
	Gunary 抛物线	$x1 / (b1 + b2 * x2 + b3 * \sqrt{x2})$ $b1 + b2 * x1 + b3 * x2^{**2}$
4	MMF	$(b1 + b2 * x1^{**} b4) / (b3 + x2^{**} b4)$
	Rational	$(b1 + b2 * x1) / (1 + b3 * x2 + b4 * x2^{**2})$
5	Gauss	$b1 * \exp (((x1 - b2)^{**2}) / b3 + ((x2 - b4)^{**2}) / b5$

备注：这些方程是精心编排的，但是，没有对每个个案都检查所使用参考文献是否正确。

对于给定的回归方程，无法给定可以反映曲线典型走向的图。函数主要是取决于各个参数的值域，从而使参数（如正负号）的变化可能产生完全不一样的曲线，使得看到的曲线起到误导而不是帮助的作用。

2.3 多元线性回归：多重共线性和其他难点

线性回归分析可以从简单双变量回归（第 2.1 节）扩展到多变量回归，也就是多元回归。在双变量个案（简单线性回归）中，从自变量 x 推导出因变量 y 。在多元回归中，通过多个自变量 x_1, x_2, \dots 的线性组合预测因变量 y 的值。在简单线性回归中，利用方程 $y = a + b * x + u_i$ （对于 y 的预测值： $\hat{y} = a + b * x$ ）解释自变量和因变量之间的关联；而在多元线性回归中，根据下面的线性方程将多个自变量引入模型： $y = a + b_1 * x_{1i} + b_2 * x_{2i} + \dots + b_n * x_{ni} + u_i$ （对于 y 的估计值： $\hat{y} = a + b_1 * x_{1i} + b_2 * x_{2i} + \dots + b_n * x_{ni}$ ）。在这里采用了残差 $e_i = y - \hat{y}$ ，以便对模型适用性进行评估（参见第 2.1.1 节“确定回归直线”和第 2.3.2 节“语法和解释 - 知识巩固”）。

y 指的是因变量。 x_1, \dots, x_n 指的是模型中的各个自变量； b_1, \dots, b_n 指的是其各自（未标准化的）影响权重（回归系数）。 $i=1, 2, 3, \dots, n$ 指的是各自自变量和影响权重（ β ）的数量。 a 是 y 轴截距或交点（又称常量 b_0 ）， u_i 是随机扰动项， e_i 是残差。在多元回归时，最小二乘法的目的是使残差的平方和最小。因此，测定的 R^2 就是衡量因变量和两个或多个自变

量最佳权重组合之间关联的量度。

多元回归假设了一个多变量因果模型（如“ x_1 、 x_2 、 x_3 造成 y ”），从而不仅可以说明多个变量是否相关或者在多大程度上相关，而且还能检验关联的方向，也就是因果模型自身的方向，以及多个自变量（又称 UV、回归量、预测变量、解释变量和影响变量）在多大程度上影响一个因变量（又称 AV、回归应变数、标准、目标变量和反应变量）。

多元回归分析的目的在于，如果相应的预测变量增大一个单位，而与此同时所有其他预测变量可以保持不变，就测定出一个回归方程，在这个回归方程中可以将回归系数解释为衡量因变量变化的量度（参见对于多重共线性的下列运算）。

下列问题是多元回归的应用实例（参见 Cohen 等人著作，2003³；Chatterjee & Price，1995²；Pedhazur，1982²）。

- 哪些参数影响产品的销量？
- 哪些因素影响上级的管理质量？
- 根据哪些实验室参数可以描述患者的康复期？

与简单回归类似，在这些问题上可以补充一些对所预测影响的方向（正向、负向）和程度的考虑。此外，例如多元回归可以调查哪些预测变量比其他的预测变量好；当进入或者剔除了一个预测变量时，预测建模在多大程度上发生变化；或者某些预测变量群组是否比其他的预测变量群组好；预测方程的优度在经过交叉验证后在多大程度上可以视为稳定。

根据因变量的尺度水平不同，使用的方法有（二项、多项）逻辑回归、有序回归，或者根据分析模型不同，可以使用路径分析、判别分析和（多变量）方差/协方差分析。

对于具有时间相依的或者季节性的结构和影响的纵向数据，可以使用 OLS 回归（参见 Woolridge 的著作，2003，第 10 和 11 章；Cohen 等人的著作，2003³，第 15 章；Chatterjee 和 Price 的著作，1995²，第 7 章）。此外，时间序列分析的专用方法可能也是比较有用的（参见 Hartung 的著作，1999，第 7 章；Schlittgen，2001；Schlittgen & Streitberg，2001⁹；Yaffee & McGee，2000）。

2.3.1 多元回归的特点

模型含有多个自变量，而非仅有一个，这就使得模型在自变量相互之间关系方面具有以下特点。

第一个特点涉及变量选择的方式，也就是从理论的角度注重内容还是从统计学的角度注重形式。由于必须有几个自变量同时起作用，因此，在根据理论推导进行建模的阶段就应保证，具有最关键内容并且测量效果最佳的变量构成一个模型，而且这个模型可以经受住从理论上对其内容的严格检验。在建模时，理论起到关键作用。不同的理论决定了不同的（等值的、相互竞争的）模型，而这些模型可以对各个预测变量做出不同的证明。缺少有关变量、出现无关变量、或者实际上呈现出曲线关联而不是假设的线性关联，这些就是通常所说的“设定误差”（参见 Pedhazur，1982²，225-230，251-254）。

与注重内容、从理论角度的建模紧密联系的，是注重形式、从统计学角度的变量选择。回

归分析提供了多种方法和标准，用于从预选的、实际上的或者潜在的有关变量（必要时从统计的预测关联性和理论的解释力方面）中选择形式上的有关预测变量。

如果待检验的模型已经确定，并且预测变量的数量和预测效果是已知的，或者至少是预设定的，则对这个模型进行回归分析时，可以选择“直接法”。

如果待检验的模型还没有确定（只有一些预选的、可能有关的预测变量，根据这些预测变量来测定模型），则为了注重形式的、从统计学角度的建模，可以使用逐步法。

在对变量集进行逐块检验（例如，分别带有计量经济学指数、教育量度、心理学实验数据的变量集）时，则所有变量进入模型。由于一个变量块中的各个变量通常是相互相关的，因此对变量逐块检验的方法就有一个风险，即某些变量进入模型时可能受到了变量块顺序的影响。相对于晚一些被引入分析的变量块的变量而言，早一些被引入分析的变量块的变量可能更容易进入模型。不同的变量块顺序可能得出不同的结论。因此，应根据估计达到的预测性能来编排变量块的顺序（Pedhazur, 1982², 164-167）。

在通过抽象的前期工作使用直接法时，应根据理论推导预先进行建模（主要是在对象拟合性、逻辑和因果性方面）。在使用逐步法时，则是由试验负责人预选变量而进行建模。因此，对于逐步测定的模型应事后检验其内容的合理性，因为逐步法不能在建模时考虑到内容方面的标准。

选择变量的工作比由统计程序来猜测要难得多。在这里一定要同意 Samprit Chatterjee 和 Bertram Price 的观点，他们在著作中两次写道，建模所依据的“变量选择工作不仅是科学，而且还是一门艺术，需要特别仔细和努力”（Chatterjee & Price, 1995², 265 和 267 页）。

自变量相互之间在统计上的无关性是进行多元回归的另一个关于形式的前提条件，因此不允许出现相互关联，更准确地说：解释变量相互之间不允许有多重共线性（可以通过 VIF/允差、特征值、条件数量和方差分量予以检验）。如果没有相互关联，则变量是相互正交的；如果有多重共线性，则变量不是相互正交的。多重共线性首先意味着，在模型中预测变量（自变量）是线性和高度相互关联的（参见 1982², 232-247）。因此，每个预测变量都决定了其他预测变量。换言之：每个自变量都可能是其他自变量的线性函数。在极端情况下，每个预测变量都可以被其他预测变量替换。

这意味着什么呢？多元回归分析的目的是测定一个回归方程，在这个回归方程中，如果相应的预测变量升高一个单位，而同时所有其他预测变量保持不变，则回归系数可以解释为衡量自变量变化的量度。但是在具有多重共线性，也就是相互高度关联时，就无法再使一个变量发生变化，而同时其他所有变量保持不变。多重共线性回归方程的回归系数、标准误差、显著性检验和置信区间不再可以解释，回归系数在输出时甚至可以带有与预计相反的正负号。无法再推导出自变量是如何影响因变量的信息，只有 R^2 是唯一可用的量度（Pedhazur, 1982², 235）。

在多元回归中利用 R^2 比较不同的模型时应注意， R^2 随着变量数量的增加而升高，而随着 N 越来越大又重新使 R^2 降低。因此，带有很少预测变量和很多个案的模型中的高 R^2 值，与带有很多预测变量和很少个案的模型的高 R^2 值具有完全不同的意义。在多元线性回归的个案中， R^2 等于观察标准和预测标准之间关联的平方（ $R^2 = r^2_{y\hat{y}}$ ）。为了比较不同的模型，必要时应使用调整的 R^2 ，与标准 R^2 相比，其优势在于不受因变量单位的影响。

因此, Chatterjee 和 Price (1995², 184, 248, 253, 259-260) 建议, 当呈现出多重共线性时, 对所有重要的、基于回归分析的结论都要十分小心 (与此成对比的是第二个例子中演示的错误解释), 这主要是因为变量选择阶段可能就已经受到了多重共线性的不利影响。例如, 对于共线性数据, 除了后退法之外, 不建议使用逐步法, 必要时可以用偏回归或者岭回归作为辅助方法, 进行带有共线性变量的建模。

可以通过 VIF (允差)、特征值、条件数量和方差分量来探索多重共线性。方差膨胀因子 (VIF) 和允差是衡量预测变量之间线性的特定量度。VIF (X_i) 是衡量各自系数可靠性的量度, 是基于预测变量 X_i 对所有其他形式为 $VIF(X_i) = 1 / (1 - R_i^2)$ 的预测变量的相关系数平方 (R_i^2)。如果预测变量之间没有线性关联, 则 $R_i^2 = 0$, 且 $VIF(X_i) = 1$ 。预测变量之间的线性关联越大, 则 R_i^2 越接近 1, VIF (X_i) 也就越大。如果 VIF 大于 10, 通常就表明有多重共线性。允差等于 VIF 的直接倒数 (允差 = $1/VIF$), 可以解释为一个变量的方差分量, 并且这个方差分量不能用模型中的其他变量解释。如果所有的预测变量相互垂直, 则 $VIF(x_i) = 1$ 。如果模型中预测变量 ($VIF(x_i) = 1 / (1 - R_i^2)$) 的所有 VIF 低于 10, 则可以认为共线性还不成什么问题。标准化预测变量的协方差矩阵 (等于非标准化预测变量的相关系数矩阵) 特征值 (又称为主分量的方差) 表明了预测变量之间是否有维度以及有多少维度。特征值小 (< 0.01) 表明存在共线性 (若特征值等于零, 则表示有完美的共线性), 同样条件数量 > 15 。条件数量是一个相关系数矩阵中的最大特征值与最小特征值之比的平方根。当条件数量 > 30 时, 应采取相应的措施 (Chatterjee & Price, 1995², 201-206 和 249)。

产生多重共线性的原因可能主要是哑变量、设定误差、抽样错误或者所调查对象的特征。在本章最后归纳了消除多重共线性的方法。

2.3.2 第一个例子: 多元回归特殊统计的解释

通过下面的多元回归, 应调查不同预测变量 (预测变量: “抱怨”、“优待”、“进修”、“业绩”、“失误”和“晋升”) 对上级行为的影响。示例和数据都摘录自 Chatterjee 和 Price (1995², 69-75) 的著作。由于所调查的模型是确定的, 因此选择直接强制进入法 (“ENTER”)。

为了计算多元回归模型, 这里建议参照 Chatterjee 和 Price (1995², 265-268) 采取的多元回归计算步骤。

语法和解释说明 (深化)

GRAPH

```
/SCATTERPLOT (MATRIX) =抱怨 优待 进修 业绩 失误 晋升
/MISSING=LISTWISE
/TITLE= '对多重共线性的第一次检验,
/FOOTNOTE= '来源: Chatterjee & Price, 19952, 70'.
```

在前面关于简单线性模型的章节已经解释了 GRAPH 语法。

REGRESSION

```
/DESCRIPTIVES MEAN STDDEV CORR SIG N
/MISSING LISTWISE
```

```

/STATISTICS COEFF OUTS CI RANOVA COLLIN TOL ZPP
/CRITERIA=PIN (.05) POUT (.10)
/NOORIGIN
/DEPENDENT 评估
/METHOD=ENTER 抱怨 优待 进修 业绩 失误 晋升
/PARTIALPLOT ALL
/SCATTERPLOT= (评估 ,*ZRESID) (评估 ,*ZPRED)
/RESIDUALS DURBIN HIST (ZRESID) NORM (ZRESID)
/SAVE ZPRED COOK LEVER ZRESID DFBETA DFFIT .

```

REGRESSION 调用用于计算线性回归分析的 SPSS 过程命令。

根据 VARIABLES = 可以列出分析的所有变量，即因变量和自变量（参见关于进口数据的示例）。简单线性回归至少需要两个变量：一个因变量，一个自变量。也可以使用“VARIABLES= (COLLECT)”来代替很长的变量表。在程序中，VARIABLES= 必须位于 /DEPENDENT 和 /METHOD 之前。

根据 /MISSING 子命令，可以给出处理可能有缺失值的方法。如果选择了 LISTWISE 选项，则整行删除所有含有缺失值的个案；如果选择了 PAIRWISE 选项，则删除所用变量含缺失值的个案。通过 MEANSUBSTITUTION 可以用变量的效应平均值代替缺失值，被代替的数值作为有效数值被引入分析（参见 Schendera 著作，2007，第 6 章）。选择了 INCLUDE，则将用户定义的缺失值作为有效数值引入分析；选项 LISTWISE、PAIRWISE 和 MEANSUBSTITUTION 可以分别添加 INCLUDE 作为补充；在选择了 MEANSUBSTITUTION 选项后，将用户定义的缺失值引入平均值的计算。一定要注意，在描述性分析和推断性统计分析中，对于缺失值的处理必须一致。

用 /DESCRIPTIVES 子命令可调用描述性统计量（MEAN，平均值、STDDEV，标准差）、皮尔逊相关系数、其单侧显著性和用于计算相关性（CORR、SIG、N）的个案数量。此外还可以调用相关系数矩阵（XPROD）平均值的方差（VARIANCE）、协方差（COV）、平方和以及叉积离差，并且只有在无法测定某些系数（BADCORR）时才显示相关系数。只有针对给定的变量和有效个案，才测定描述性参数。对于 PAIRWISE 和 MEANSUBSTITUTION 选项，描述性统计量是基于每个变量所有个案的有效数值；对于 LISTWISE 选项，则除了所有变量之外，只将带有有效数值的变量引入描述性统计量。如果给出了子命令 /ORIGIN（见下文），则在计算描述性统计量时，假设其平均值为 0。

用 /STATISTICS 子命令可以调用回归方程和自变量的统计量。回归方程的选项包括 R（多重 R，包括 R²、调整 R² 和所显示估计值的标准误差）、ANOVA（方差分析，包括回归的平方和或者离差、平均平方和、F 值和 F 值的显著性）、CHA（各个步骤之间 R² 的变化，包括 F 值和 F 值的显著性，只适用于逐步法）、BCOV（非标准化回归系数的方差-协方差矩阵）、XTX（XTX 矩阵，sweep matrix）、COLLIN（共线性诊断，其中主要是 VIF 方差膨胀因子，具有尺度的或者不居中的叉积矩阵的特征值、条件指数和方差分量的特征值）和 SELECTION（选择统计量，主要是 AIK、PC、Cp 和 SBC）。用于自变量的选项包括 COEFF（非标准化回归系数（B）、系数的标准误差、标准化回归系数（ β ）、T 和单侧 T 显著性）、OUTS（还没有进入方程的变量的统计量： β 、T、双侧 T 显著性和最小允差）、ZPP（零阶相

关、部分相关和偏相关)、CI (非标准化回归系数的 95%置信区间)、SES (标准化回归系数大约的标准误差)、TOL (回归方程中的变量的允差; 对于不在回归方程中的变量则分别测定一个允差, 就好像后面这个变量会作为唯一一个变量而进入回归方程一样) 和 F (非标准化回归系数的 F 值及其显著性)。STATISTICS 必须要在 DEPENDENT 和 METHOD 之前给定。通过 ALL 可以调用除了 F 之外的所有选项。

根据 /CRITERIA 子命令, 可以通过 PIN (适用于 FORWARD 和 STEPWISE 方法) 和 POUT (适用于 BACKWARD 和 STEPWISE 方法) 确定模型根据哪些参数进入或者剔除变量。方法 ENTER、REMOVE 和 TEST 使用 TOLERANCE 选项。通过 PIN (.05), 预设了决定一个变量是否可以进入模型的数值。如果一个变量的统计量值小于进入值, 则这个变量就进入模型; 所给出的进入值 (PIN 或者 FIN) 越大, 变量就越可能进入模型。SPSS 中预设置的 0.05 被认为是相对保守的, 为了使潜在的有关预测变量进入模型, 最多可以接受 0.20。利用 POUT (预设置 0.10) 定义衡量从模型中剔除哪一个变量的数值。如果变量的概率大于剔除值, 则将这个变量剔除; 给定的进入值 (POUT, 也可用 FOUT) 越大, 变量保留在模型中的可能性也就越大。进入值必须小于剔除值。预设置的 FIN 值为 3.84, FOUT 值为 2.71。如果给定了 PIN 和 POUT 的数值, 则在 PIN 和 POUT 下面给定的 F 值就不起作用了。PIN/POUT 和 FIN/FOUT 的作用方式并不相同。由于 F 值 (PIN/POUT) 的显著性水平取决于自由度的数量和已进入的变量的数量, 因此, 只有在留有一个预测变量的情况下, PIN/POUT 和 FIN/FOUT 才能得出相同的结果。

通过选项 TOLERANCE 可以针对 ENTER、REMOVE 和 TEST 方法给定可靠的允差, 这个允差可以具有一个变量, 从而可以被引入分析。简单地说, 一个变量的允差是其与其他自变量的可靠相关性的程度 (前面在介绍多元回归分析时, 已经探讨了多重共线性问题)。允差很小的变量差不多就是其他变量的线性函数。可以给定一个 0 到 1.0 之间的数值作为允差, 以衡量被接受的相互关联的程度。带有低于允差的数值的变量不进入模型; 如果这些变量进入模型, 将会使对模型的分析不可靠或者不稳定。通过 CIN[(数值)], 可以为计算特定的临时变量调整置信区间的百分比值。在 MAXSTEPS[(n)] 一项下, 可以针对给定的方法预先设定步骤的最大数量, 尤其是对于采用逐步法时不是最佳的进入和剔除标准, 可以通过 MAXSTEPS 避免循环不断地剔除和进入变量。预设值取决于所使用的方法, 例如, 在使用 STEPWISE (逐步法) 时, 等于自变量数量的两倍。

CRITERIA 必须要在 DEPENDENT 和 METHOD 之前给定。

子命令 /NOORIGIN 将常量项引入模型 (预设置)。如果给定了/ORIGIN, 则回归直线会穿过原点, 从而就不会输出常量。如果给定了 ORIGIN, 则在计算描述性统计量时, 就好像平均值是 0 一样。在读取和写入矩阵数据时, 每次都应使用同样的子命令。/NOORIGIN 和 /ORIGIN 必须要在 DEPENDENT 和 METHOD 之前给定。

在/DEPENDENT 子命令下必须给定回归模型的因变量, 也就是要对其测定一个回归方程的变量。例如, 在本例中就是变量“评估”。必须至少给定一个因变量。如果给定了几个因变量, 则利用相同的设置、方法、标准和自变量分别测定回归模型和回归方程。如果估计回归模型有几个因变量, 则可以参见 SPSS 过程命令 ANOVA、MANOVA 和 GLM (见第 6 章)。只能给定一个 DEPENDENT 子命令。必须至少有一个 METHOD 子命令跟在 DEPENDENT 子命

令之后。根据子命令/METHOD，可以给定选择变量的方法（在本例中是 ENTER），根据这个方法（例外情况是 REMOVE，TEST）给定自变量，在本例中就是“抱怨”、“优待”和“晋升”。

为了设定模型，可以单独或者成块地用下列方法中的一个进入或者剔除变量：“进入”（ENTER，预设置）、“步进”（STEPWISE）、“剔除”（REMOVE）、“向前”（FORWARD）、向后（BACKWARD）和 TEST。如果预测变量的数量和预测效果是已知的，或者至少是预先给定的，则应使用直接法。如果预测潜力不明或者要测定一个只含有很少变量的有效预测模型，则可以使用逐步法。形式上逐步选择的变量的顺序，并不说明其内容上的关联性（Chatterjee & Price, 1995², 252）。“步进”、“剔除”、“向前”和“向后”这四种方法在只有一个预测变量时也可以使用，但是其意义非常有限。

直接法

ENTER（进入）。在一个步骤中，所有变量进入模型（因此在步骤 1 后，ENTER 立即停止）。如果给定了几个变量，则这几个变量进入模型的顺序取决于其允差，根据允差值递减的顺序进入模型。如果预测变量的数量和预测效果是已知的，或者至少是预先给定的，则使用 ENTER。如果预测潜力不明或者要测定一个只含有很少变量的有效预测模型，则可以使用逐步法（如通过 STEP）。

REMOVE（剔除）。在一个步骤中剔除一个组块的所有变量。

TEST（子集）（子集）（…）…。在 TEST 后面的括号里给定的变量子集首先成块地一次性添加到回归方程中，然后从方程中剔除每个子集，并显示 R^2 统计量和相应的显著性。

逐步法

STEPWISE（步进）。如果在回归方程中已经含有变量，则首先剔除显著性数值最大的变量，前提是这个数值大于 PORT。在没有这个变量的情况下，重新计算回归方程，不断重复这个过程，直到无法再剔除自变量为止。然后方程中不包含的、显著性数值最小的自变量进入模型，前提是这个值小于 PIN。现在重新检验模型中的所有变量是否要剔除。不断重复这个过程，直到变量无法再剔除或者进入为止，或者达到预设置的迭代最大次数（MAXSTEPS）。

FORWARD（向前）。依次检验变量是否可以进入模型。首先是显著性数值最小的变量进入模型，前提是这个数值小于 PIN。不断重复这个过程，直到不再有变量达到进入标准为止。

BACKWARD（向后）。第一步先由模型一次性进入所有变量，然后依次检验是否可以剔除。首先剔除显著性数值最小的变量，前提是这个数值大于预设置的 POUT。不断重复这个过程，直到不再有变量达到剔除标准为止。

根据这个方法给定了预测变量，在本例中是“抱怨”、“优待”和“进修”等。必须至少给定一个预测变量。

必须要给定一个/METHOD 子命令，可以同时给定几个/METHOD 命令行。如果给定了几个 METHOD 子命令，则始终将首先给定的预测变量添加到来自后续给定 METHOD 子命令的预测变量。例如，如果在第一个/METHOD 命令中只给定了变量 VAR1，在第二个/METHOD 命令行中给定了变量 VAR2，则第二个回归方程的变量列表就是 VAR1 VAR2。

通过子命令/RESIDUALS 可以调用一个范围广阔的残差诊断（参见第 2.1.3 节）。各个不同选项用于命名很多特征量的统计分析和图形分析，主要是杜宾-瓦森、预测值、残差或者距离。例如，（在本例中没有使用的）选项 ID（变量名）对于个案诊断非常有用，在识别离群值、显示图形（/RESIDUALS、/SCATTERPLOT 和/PARTIALPLOT）中的点时，如果没有给定用户定义的 ID 变量，则可以根据有效工作数据集的行号（“个案编号”）来识别个案。DURBIN 选项调用杜宾-瓦森统计量。杜宾-瓦森统计量对于评估自相关非常重要。HIST 在本例中调用一个直方图，其中包括临时变量 ZRESID 的正态分布曲线图。ZRESID 包含了模型的标准化残差（也允许用 PRED、ZPRED、ADJPRED、SEPPRED、RESID、DRESID、SRESID、SDRESID、MAHAL、COOK 和 LEVER）。通过 NORM 或 NORMPROB 选项，为 ZRESID 调用一个 P-P 图（也允许用 PRED、ZPRED、RESID、DRESID、SRESID 和 SDRESID）。始终标准化地显示 P-P 图，也就是说，调用 PRED、RESID 或者 DRESID 就会标准化地显示 ZPRED、ZRESID 或者 SDRESID。

更详细的内容，请参见 Cohen 等人（2003³）的著作。细节请参见前文简单线性回归的引言部分。

预测值选项详述

回归模型中预测值的选项如下。

PRED: 非标准化预测（估计）值。由模型来预测因变量的数值。

ZPRED: 标准化预测值。在标准化一个预测值（估计值）时，将预测值与平均预测值的差除以预测值的标准差。标准化预测值（估计值）的平均值为 0，标准差为 1。

ADJPRED: 调整预测值。是指一个被从回归系数的计算中删除的个案的预测值。

SEPPRED: 预测值的标准误差。是指一些个案的因变量平均值的标准差估计值，并且这些个案与因变量具有同样的数值。

测定的杠杆值选项详述

杠杆值指的是一个个案对于估计值确定过程的影响。如果一个个案（离群值）从空间上距离分布图形其余部分的中心（平均值、矩心）很远，则这个个案就有很大的杠杆作用。杠杆量度包括马氏距离和杠杆值（参见 Cohen 等人的著作，2003³，395）。

LEVER: 衡量一个点对回归拟合优度影响的杠杆值。杠杆作用的居中值在 0（对拟合优度无影响）和 $(N-1)/N$ 之间波动。

MAHAL: 马氏距离确定了一个个案的自变量数值与所有个案的平均值相差有多大。因此，较大的马氏距离就表明一个个案的一个或者多个自变量具有极高的数值。

残差（差异）选项详述

残差指的是因变量观察值和相应的预测值之间的距离。在 Cohen 等人的著作（2003³，398）中，称残差为差异统计量。

RESID: 非标准化残差。观察值和根据模型的预测值之间的差值。

ZRESID: 标准化残差。残差除以其一个标准误差估计值的商。标准化残差，也就是我们所说的皮尔逊残差的平均值为 0，标准差为 1。

DRESID: 学生化残差。是指一个被除以其估计的标准差的残差，并且根据个案的因变量数值与自变量平均值之间的差值不同，每个个案的这个标准差也是不同的。

SRESID: 删除残差。指的是没有引入回归系数计算的一个个案的残差。也就是因变量数值和调整估计值之间的差值。

SDRESID: 学生化删除残差。一个个案的删除残差除以其标准误差的商。学生化删除残差和相应的学生化残差之间的差值说明了，在删除一个个案前后这个个案的预测有多大区别。

测定的影响量度选项详述

影响量度是杠杆作用和差异的后果。作为一种量度，高杠杆值衡量的是，如果从回归函数的测定中删除了某个个案，则所有其他个案的残差发生了多大变化。通常，每次只从估计过程中删除一个个案（参见 Cohen 等人著作，2003³，402）。两个影响量度 DfFit 和 Cook 距离基本上是等价的，唯一的区别是 Cook 值不是负值。

影响量度 DfFit（“difference in fit, standardized”的缩写）指的是由于删除某个观察值而产生的预测值变化。对于标准化 DfFit 适用的规范是，除以平方根 p/N 后检验所有绝对值大于 2 的个案，其中 p 是方程中自变量的数量， N 是个案的数量。Cohen 等人（2003³，404）建议，对于中等或者较大的数据量使用截断点 1 或 2。Chatterjee 和 Price（1995²，89）建议，针对这个影响量度，检验所有明显很高的数值。

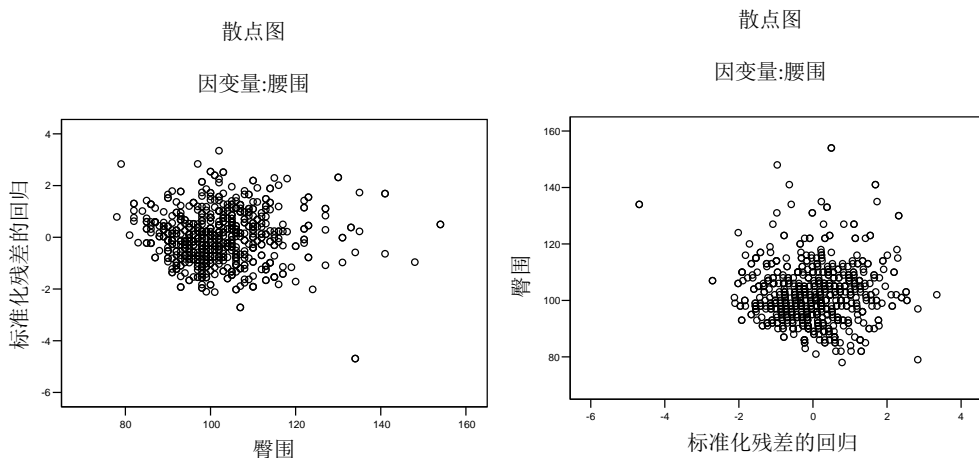
针对与预测模型中的均方误差成比例的数据集，Cook 距离指的是整个数据集的估计值与减少了数据集中的某一个观察值后的估计值之间的均方误差。较高的 Cook 距离表明，删除相应的个案对其余个案回归函数的计算产生了很大影响。根据通行的规范，Cook 距离超过 1 就视为具有高杠杆（参见 Cohen 等人著作，2003³，404）。Chatterjee 和 Price 在著作（1995²，89）中建议，对所有明显高的 Cook 距离进行检验。

利用 /SCATTERPLOT 子命令调用三个双变量散点图，由系统设定的关键词在这里被加上前缀*，以使其区别于用户设定的变量。允许使用关键词 PRED、RESID、ZPRED、ZRESID、DRESID、ADJPRED、SRESID、SDRESID、SEPPRED、MAHAL、COOK 和 LEVER。注意：需要用另一个名称将这些变量存储到数据集中（参考 GRAPH 命令）。为了创建一个散点图，在模型中必须至少含有一个自变量。与鼠标控制相反，通过 /SCATTERPLOT 子命令的语句也可以给定自变量。

SPSS 在这里的说明可能一开始让人不太容易理解。在输出的 Scatterplot 中的标题始终指的是模型的因变量，而不是图形中的因变量。例如，如果自变量（如下面的例子所示）定位到（在 Y 轴上）因变量的位置，则通过在图形中自动分配的标题将其命名为看起来是因变量，这起初可能就让人感到迷惑。在左图中，因变量是在 Y 轴上截取的。在右图中混淆了 x 变量和 y 变量（从 X 轴和 Y 轴的正确名称中可以看出）。但是自动分配的标题只是看起来将 Y 轴命名为因变量。因此，这个标题始终是指模型的变量，而不是图形 Y 轴上的变量。

例如

```
...
/SCATTERPLOT= (*ZRESID, huefte)
/SCATTERPLOT= (huefte, *ZRESID)
...
```



类似情况也适用于子命令/PARTIALPLOT 或 PARTIAL 绘制的偏残差图。偏残差图是基于自变量和因变量的残差，前提是这两种变量针对剩余的自变量分别进行一次回归。为了创建一个偏残差图，在模型中必须至少含有两个自变量。

如果通过/SELECT 命令调用了一个交叉验证，则可以根据（预设置的）SEPARATE 选项设定，针对所选择的和未选择的个案分别进行一次残差诊断。相反，POOLED 命令归纳了残差诊断的所有个案。

通过/CASEWISE 子命令，可以针对残差调用内容丰富的个案诊断。通过 ALL 命令将个案诊断中的所有个案输出；通过 OUTLIERS（数值）可以针对范围广阔的数据集将输出结果限定带有标准化残差的个案，并且这些残差的标准差超过预设值（预设置：3）（也就是“离群值”）。

ALL 和 OUTLIERS 命令的可用关键词有 DEPENDENT（因变量）、PRED、RESID、DRESID、SRESID、SDRESID、MAHAL、LEVER 和 COOK。所有这些变量可以一次性调用；预设置的是 DEPENDENT、PRED 和 RESID。如果通过 ID 给定了一个 ID 变量，则在输出结果中除了当前的行号之外还分别列出了 ID 数值，从而十分方便个案的识别。这是由于这些 ID 数值不受数据组临时分类的影响。如果给定了选项 ALL，则忽略选项 OUTLIERS（数值）。如果既没有给定 ALL 也没有给定 OUTLIERS，则 SPSS 将 OUTLIERS 视为预设置。

通过 PLOT（ZRESID；预设置）针对一个个案图形调用所给定残差类型的输出结果（也可以使用 RESID、DRESID、SRESID、SDRESID）。在个案图形中只能展示残差，并且每次只能展示一个残差类型。如果调用的话，则在输出结果中显示标准化的 RESID 和学生化的 DRESID。

可惜无法给定多个 /CASEWISE 命令行。

通过子命令/SAVE，可以将多个残差和影响统计量保存到当前的工作数据集中。除了前面已经介绍的预测值、杠杆值和残差（差异）之外，还可以保存影响统计量，即 DfBeta、标准化 DfBeta、DfFit、标准化 DfFit 和协方差分量。

DFBETA：非标准化 DfBeta。贝塔（ β ）值中的差值产生的原因，是指由于删除某个特定个案而导致回归系数发生变化。针对模型中的每一项，包括常量在内，都计算出一个数值。

SDBETA：标准化 DfBeta。贝塔（ β ）值中的标准化差值。是指由于删除某个特定个案而导致的回归系数发生变化。建议将绝对值大于 2 的个案除以 N 的平方根，以对其进行检验（在这里 N 是个案的数量）。针对模型中的每一项，包括常量在内，都计算出一个数值。

DFFIT：DfFit。是指由于删除某个特定个案而导致的预测值变化。

SDFIT：标准化 DfFit。拟合值中的标准化差值。由于删除某个特定个案而导致估计值发生变化。建议将所有绝对值大于 2 的个案除以平方根 p/N 以做检验，其中 p 是方程中自变量的数量， N 是个案的数量。

COVRATIO：协方差比例。是指在从回归系数计算中剔除某个个案时的协方差矩阵行列式与所有个案进入回归系数计算时的协方差矩阵行列式的比例。如果两者之比接近 1，则剔除的个案对协方差矩阵的影响很小。

FITS：一次性调用 DFBETA、SDBETA、DFFIT、SDFIT 和 COVRATIO。

根据存储方式不同，在数据集里面可以存储用户定义的和系统定义的变量名。

系统定义的变量名示例：

/SAVE=COOK LEVER

在活动数据集中，将数值保存为自动分配了名称 COO_1 和 LEV_1 的变量。

用户定义的变量名示例：

/SAVE COOK (CWERTE) LEVER (LWERTE)

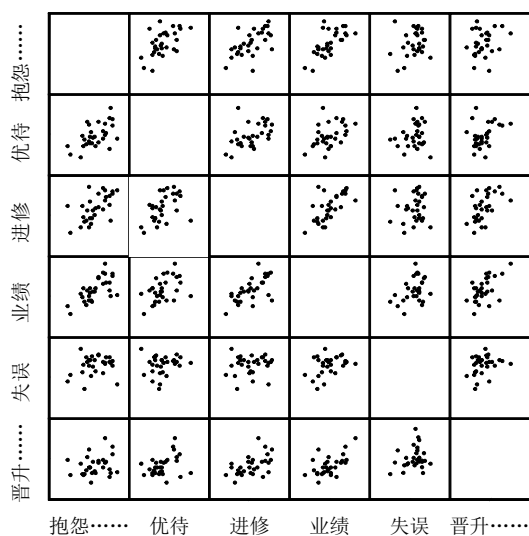
在有效的数据集中，将同样数值保存为由用户分配了名称“CWERTE”和“LWERTE”的变量。

输出结果

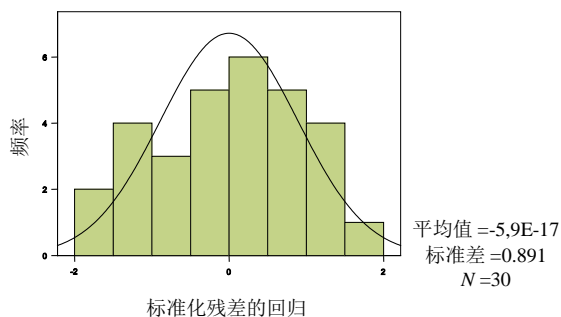
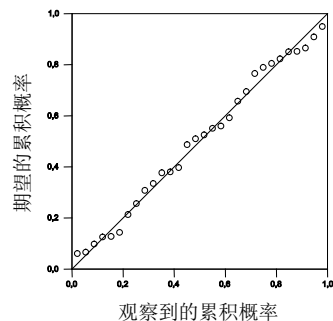
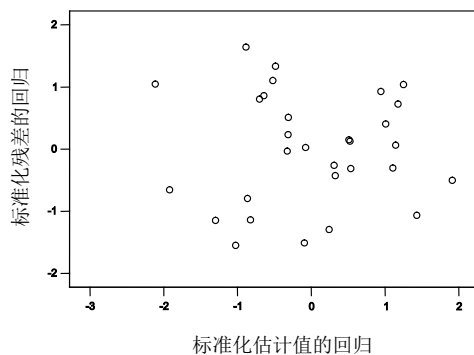
与简单线性回归类似，在下面的输出结果中把图形式残差分析写在前面，以便在检查所测定的统计量之前可以评估模型的拟合程度。由于在关于双变量线性回归的一章中已经介绍了其中的很多输出元素，因此本节只重点介绍多元回归的数据、模型以及统计量的特点。基础知识请参见关于双变量线性回归一章。

预测变量中的线性测量值分布可能初步说明了存在多重共线性，这是多元回归的特有问题。双变量散点图绝不会明显地含有线性分布。如果不是多个点散布地非常广泛，那么一定程度上的线性表明了“进修”和“成绩”之间的关联。多重共线性的可能性不大。这些分布的非线性也表明了，之后测定的皮尔逊相关系数是无效的，因为这个系数是以（不存在的）线性为前提条件的。

对多重共线性的第一次检验

来源: Chatterjee & Price, 1995², 70

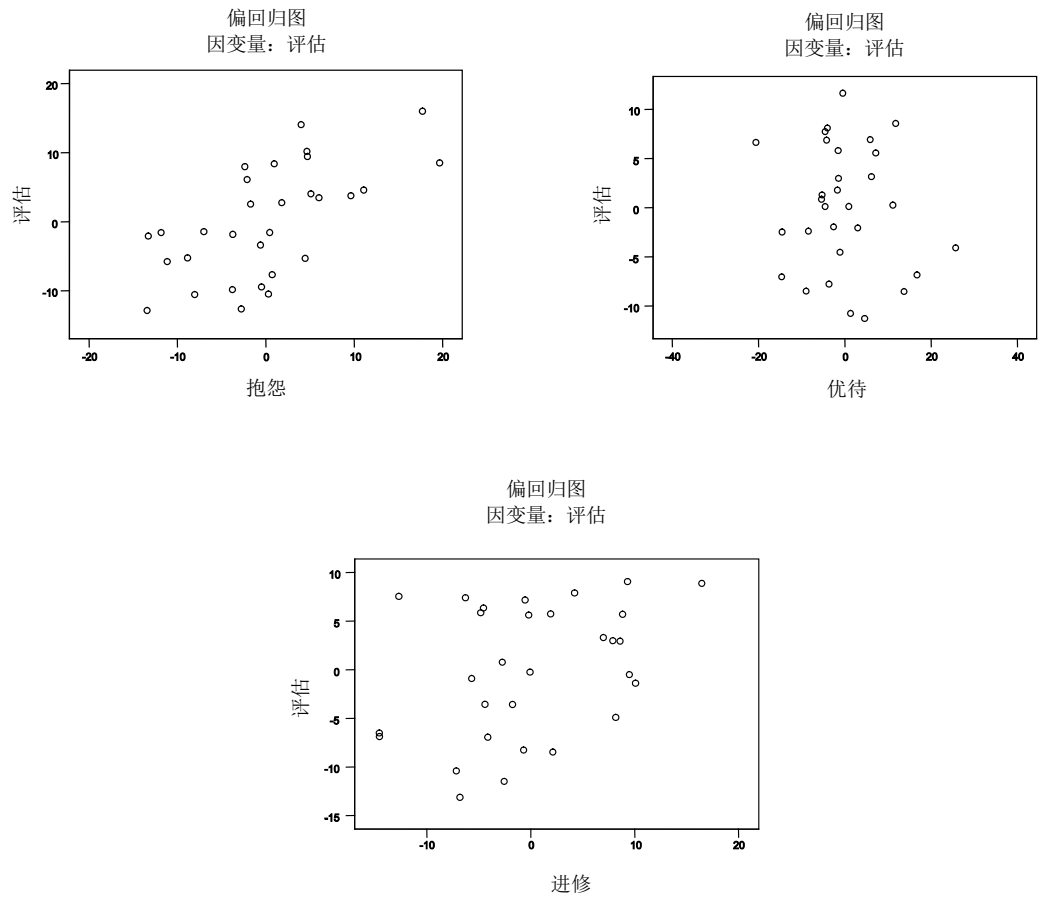
图形式残差分析:

直方图
因变量: 评估标准化残差的 P-P 图
因变量: 评估散点图
因变量: 评估

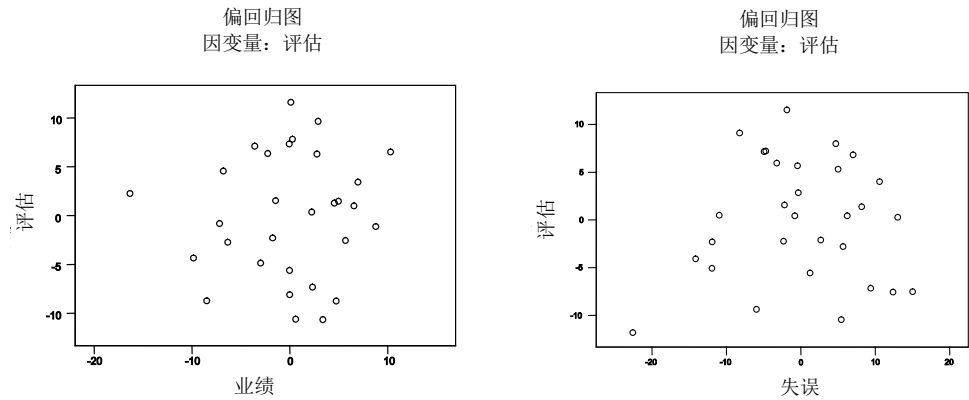
在直方图中没有发现离群值。在 P-P 图中, 所有标准化残差都在正态分布的基准线上(线性, 正态性)。在根据预测值的标准化残差散点图中, 只发现残差呈现所期望的随机散布, 但

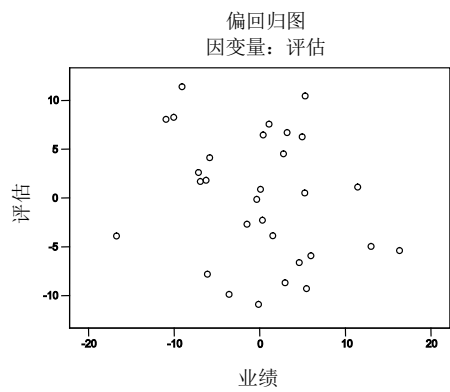
没有发现（不期望的）系统性散布。因此，没有迹象表明存在非线性关联或者模型误设定。

在多元回归中，针对模型中含有的每个预测变量都调用一个单独的、表现所测定残差的散点图。数据在这些偏回归图中应呈随机分布。



在表示“抱怨”、“优待”和“进修”的偏回归图中，没有迹象表明模型假设已经不成立。





同样，在表示“业绩”、“失误”和“晋升”的偏回归图中，也没有迹象表明模型假设不成立。

尽管通过选项 CASEWISE 调用了残差的详细列表，但是 SPSS 针对这个模型不输出“个案诊断”表。原因是：这个模型不含符合预设定义所定义的离群值，这同样表明模型拟合是成功的。“残差统计量”表中的最大值和最小值证实了这个诊断结果。

残差统计量 ^a					
	最小值	最大值	平均值	标准差	N
非标准化预测值	42.60	84.55	64.63	10.419	30
标准化预测值	-2.114	1.911	0.000	1.000	30
预测值的标准误差	1.635	5.209	3.269	1.003	30
调整预测值	40.32	88.68	64.99	10.736	30
非标准化残差	-10.942	11.599	0.000	6.294	30
标准化残差	-1.548	1.641	0.000	0.891	30
学生化残差	-1.861	1.769	0.021	1.009	30
删除残差	-15.818	13.479	0.361	8.184	30
学生化剔除残差	-1.975	1.861	0.024	1.031	30
马氏距离	0.585	14.787	5.800	4.176	30
Cook 距离	0.000	0.221	0.045	0.051	30
居中杠杆值	0.020	0.510	0.200	0.144	30

a. 因变量：评估。

例如，在“残差统计量”表格中，居中杠杆值非常低，Cook 距离远远小于 1。同样，标准化残差小于 2.0，但值得注意的是，残差的可用性仅限于简单（双变量）回归模型，不适用于检验多元回归（Chatterjee & Price, 1995², 86）。

多元回归分析的统计量

描述性统计量			
	平均值	标准差	N
评估	64.63	12.173	30
抱怨	66.60	13.315	30
优待	53.13	12.235	30

续表

	平均值	标准差	N
进修	56.37	11.737	30
业绩	64.63	10.397	30
失误	74.77	9.895	30
晋升	42.93	10.289	30

表“描述性统计量”反映了个案的平均值、标准差和数量。在多元回归的个案中，不能直接用“模型总结”表调整所给定的标准差。

相关性

	评估	抱怨	优待	进修	业绩	失误	晋升
皮尔逊相关系数	评估	1.000	0.825	0.426	0.624	0.590	0.156
	晋升	0.825	1.000	0.558	0.597	0.669	0.188
	抱怨	0.426	0.558	1.000	0.493	0.445	0.147
	失误	0.624	0.597	0.493	1.000	0.640	0.116
	优待	0.590	0.669	0.445	0.640	1.000	0.377
	进修	0.156	0.188	0.147	0.116	0.377	1.000
	业绩	0.155	0.225	0.343	0.532	0.574	0.283
显著性（单侧）	评估		0.000	0.009	0.000	0.000	0.205
	晋升	0.000		0.001	0.000	0.000	0.160
	抱怨	0.009	0.001		0.003	0.007	0.219
	失误	0.000	0.000	0.003		0.000	0.271
	优待	0.000	0.000	0.007	0.000		0.020
	进修	0.205	0.160	0.219	0.271	0.020	
	业绩	0.207	0.116	0.032	0.001	0.000	0.065
N	评估	30	30	30	30	30	30
	晋升	30	30	30	30	30	30
	抱怨	30	30	30	30	30	30
	失误	30	30	30	30	30	30
	优待	30	30	30	30	30	30
	进修	30	30	30	30	30	30
	业绩	30	30	30	30	30	30

表“相关性”含有两条信息：（1）所有预测变量相互相关（这不是我们想要的；残差的目标是希望：相关性最小化），（2）预测变量与因变量相关（这个是我们想要的；残差的目标是希望：相关性最小化）。仅仅检查皮尔逊统计量的数字是不够的，必须通过图形进行前提条件检验，得出线性的前提条件（参见上文）。如果是非线性关联，则皮尔逊相关系数无效。只有当图形中呈现线性分布，并且显著性 $p < 0.05$ （单侧）时，才能得出双变量线性关联的结论。

进入/剔除的变量^b

模型	进入的变量	剔除的变量	方法
1	晋升 抱怨 失误 优待 进修 业绩 ^a		进入

a. 进入了所有想要的变量。

b. 因变量：评估。

表“进入/剔除的变量”展示了在每个步骤进入的、或者根据方法不同重新剔除的预测变量。例如，如果这个方法不剔除变量或者所有进入的变量都符合预设置的标准，则“剔除的变量”一列保留空白。从表下面的说明可以看出，所有想要的变量都已经进入，“评估”是模型的因变量。

模型总结^b

模型	R	R ²	调整 R ²	估计值的标准误差	杜宾-瓦森统计量
1	0.856 ^a	0.733	0.663	7.068	1.795

a. 预测变量：（常量），晋升、抱怨、失误、优待、进修和业绩。

b. 因变量：评估。

从表“模型总结”中可以获取针对每个模型或者每个步骤汇总的、关于相应模型与因变量之间关联的量度。对于直接法，通常只输出一行。

对于多元回归模型，尤其要仔细观察 R、调整 R²、预测变量的标准误差和杜宾-瓦森统计量。R 表示因变量的观察值和模型预测值之间线性相关的程度（极差：0~1）。如果有多个预测变量，则 R 等于多重相关系数；如果模型中只有一个预测变量，则 R 等于皮尔逊相关系数。而且对于多元回归适用这个解释：R 越高，模型和因变量之间的关联就越大。R = 0.856 表示模型和因变量之间存在很大的关联。R²（决定系数，R²）= 0.733 表明模型可以解释因变量中大约 3/4 的变异。模型适合解释多变量关联的影响。杜宾-瓦森统计量检验残差是否自相关。测定值 1.795 尽管表明了正自相关，但是不能直接得出残差独立性具有显著性的结论。只有在根据杜宾-瓦森表（T=30，K=6，α=0.05）进行检验后才可以确定，测定值尽管指向正自相关的方向，但是还是在无关带范围内（L_U = 1.07060，L_O = 1.83259）。可以排除自相关的可能性。对于含有多个预测变量的模型应注意，R²也取决于回归量的数量。在比较多个模型时，使用“调整 R²”这个量度是很有必要的。“调整 R²”是根据自变量的数量做了调整。

ANOVA^b

模型		平方和	df	平方均值	F	显著性
1	回归	3147.966	6	524.661	10.502	0.000 ^a
	残差	1149.000	23	49.957		
	总值	4296.967	29			

a. 预测变量：（常量），晋升、抱怨、失误、优待、进修和业绩。

b. 因变量：评估。

方差分析是借助于 F 检验来检验模型的回归解释。其中的 F 值是根据回归（解释方差）与残差（误差方差）的比例得出的。正如上面例子所示，如果模型达到了统计显著性，则说明自变量很好地解释了因变量的变异；但是，如果得到的显著性超过预设定的 α ，则模型不能解释因变量的变异。

系数^a

模型	非标准化系数		标准化系数	T	显著性	相关性			共线性统计量	
	B	标准误差	β			零阶	相关	部分相关	允差	VIF
1 (常量)	10.787	11.589		0.931	0.362					
抱怨	0.613	0.161	0.671	3.809	0.001	0.825	0.622	0.411	0.375	2.667
优待	-0.073	0.136	-0.073	-0.538	0.596	0.426	-0.112	-0.058	0.625	1.601
进修	0.320	0.169	0.309	1.901	0.070	0.624	0.368	0.205	0.440	2.271
业绩	0.082	0.221	0.070	0.369	0.715	0.590	0.077	0.040	0.325	3.078
失误	0.038	0.147	0.031	0.261	0.796	0.156	0.054	0.028	0.814	1.228
晋升	-0.217	0.178	-0.183	-1.218	0.236	0.155	-0.246	-0.131	0.512	1.952

a. 因变量：评估。

为了使表格能完整地在这一页呈现，利用上述语句调用的置信区间（选项“CI”）已经被删除。

对于回归系数的解释，重要的是显著性、T 值、回归系数值和正负号。通常，只解释显著的预测变量，如“抱怨”。在解释标准化（非标准化）系数时应注意其各自的优点和缺点（为此参见双变量例子中的引言部分）。利用在表“系数”中输出的、显著性的非标准化系数，可以这样表述回归方程：“评估”=10.787+0.613*“抱怨”。

（备注：也含有非显著性变量的回归方程，例如：

“评估”=10.787+0.613*“抱怨” - 0.073*“优待” + 0.320*“进修” + 0.082*“业绩” + 0.038*“失误” - 0.217*“晋升”，

将会得出另一个结果。

标准化回归系数（0.671）表明了两件事：影响相对较高（最大值为 1），并且影响方向为正。若“抱怨”值高，则相应的“评估”值高；若“抱怨”值低，则相应的“评估”值低。

“零阶”相关性等于表“系数”中的双变量相关系数，指的是因变量和各自预测变量之间的线性关联。例如，在本例中，“抱怨”和“评估”之间的零阶皮尔逊相关系数为 0.825（参见表“系数”）。“偏”一列含有偏相关系数，“部分”一列含有半偏相关系数（参见 Cohen 等人著作，2003³，69-75）。

在偏相关系数中，无论是根据预测变量，还是根据皮尔逊相关系数中的因变量，都可以计算出所有其他预测变量的预测（其他预测变量保持“恒定”）。例如，只要从“抱怨”和“评估”中剔除其他预测变量的影响，则这两个变量之间的相关性就下降到 0.622。“部分”一列含有半偏相关系数；在这里，只根据相关的那个预测变量计算出了所有其他预测变量的影响。例如，只要从“抱怨”中剔除了所有其他预测变量的影响，则“抱怨”和“评估”之间的相关

性就下降到 0.411，从而表明了一个预测变量对因变量的“纯”影响。因此，对于估计一个预测变量的影响而言，半偏相关系数是一个十分重要的量度。即使预测变量不是相互关联的，半偏相关系数加起来也不必达到 R^2 （在本例中加起来达到 0.763； $R^2 = 0.733$ ）。

“VIF”和“允差”是多重共线性的量度。模型中预测变量的所有 VIF 都低于 10，因此不存在共线性。只有当 VIF 很高，也就是有多重共线性的迹象时，才应查看额外输出的“多重共线性诊断”表，以便能够更准确地判断多重共线性的类型。

多重共线性诊断^a

模型	维度	特征值	条件指数	方差分量						
				(常量)	抱怨	优待	进修	业绩	失误	晋升
1	1	6.875	1.000	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2	0.040	13.132	0.00	0.08	0.20	0.00	0.00	0.02	0.31
	3	0.035	13.968	0.08	0.00	0.06	0.07	0.00	0.13	0.18
	4	0.024	17.104	0.00	0.09	0.66	0.16	0.02	0.01	0.03
	5	0.013	23.208	0.14	0.25	0.01	0.61	0.08	0.00	0.12
	6	0.008	30.118	0.75	0.08	0.02	0.15	0.02	0.68	0.12
	7	0.006	34.409	0.02	0.50	0.05	0.01	0.88	0.16	0.24

a. 因变量：评估。

在下一节，将根据一个关于多重共线性的例子介绍“多重共线性诊断”表的内容及其解释。

综上所述，没有迹象表明测定的模型和变量选择可能受到了多重共线性的不利影响。

2.3.3 第二个例子：多重共线性的识别和消除

下一个例子调查了计量经济学数据，也就是从 1949 年至 1959 年 ($n = 11$) 从法国进口的数据，达到数十亿法郎之巨。纽约的 Samprit Chatterjee 教授友好地提供了这个（纽约）例子和相关数据。

这个例子说明了关于自变量和因变量之间预计关联的理论和假设的作用。人们认为，变量 INPROD（国内生产总值）、KONSUM（消费）和 LAGER（库存）可以用来描述或者预测 IMPORTE（进口额）的规模。假设之一是，提高国内生产总值也会导致进口额增加（如原材料）。因此，预计在 INPROD 和 IMPORTE 之间存在正相关。对于建模的其他方面，请参见 Chatterjee & Price（1995²）的例子。

Chatterjee & Price（1995²，197-198，228-235-235）建议在预测变量的数量明确时，对所有可能的预测变量组合分别进行回归分析，例如，在本例中：

模型 1：IMPORTE = INPROD

模型 2：IMPORTE = LAGER

模型 3：IMPORTE = KONSUM

模型 4: IMPORTE = INPROD LAGER
模型 5: IMPORTE = INPROD KONSUM
模型 6: IMPORTE = LAGER KONSUM
模型 7: IMPORTE = INPROD LAGER KONSUM

对于所测定的系数，下列现象可以表明具有多重共线性：

- 当添加或者减少一个预测变量时系数发生较大变化。
- 当删除或者转换一个个案时系数发生较大变化。
- 系数的正负号与理论或期望不符。
- 有关变量的系数有较大的标准误差。

由于篇幅所限，这里没有详细阐述这种操作方法，但是强烈建议在分析实践中予以采用。
当变量数量较多时，偏回归或者岭回归在选择变量方面或许是很有帮助的（见第 5 章）。

语句（没有特别说明）

只利用所选择行的进口数据（ $n = 11$ ）进行多重共线性计算。

```
data list list
/x (a)      行    年份  进口额          国内生产总值      库存      消费额
begin data
X           1      49   15.9          149.3          4.2      108.1
X           2      50   16.4          161.2          4.1      114.8
X           3      51   19.0          171.5          3.1      123.2
X           4      52   19.1          175.5          3.1      126.9
X           5      53   18.8          180.8          1.1      132.1
X           6      54   20.4          190.7          2.2      137.7
X           7      55   22.7          202.1          2.1      146.0
X           8      56   26.5          212.4          5.6      154.1
X           9      57   28.1          226.1          5.0      162.3
X          10      58   27.6          231.9          5.1      164.3
X          11      59   26.3          239.0          0.7      167.6
X          12      60   31.1          258.0          5.6      176.8
X          13      61   33.3          269.8          3.9      186.6
X          14      62   37.0          288.4          3.1      199.7
X          15      63   43.3          304.5          4.6      213.9
X          16      64   49.0          323.4          7.0      223.8
X          17      65   50.3          336.8          1.2      232.0
X          18      66   56.6          353.9          4.5      242.9
end data.
compute Jahr=Jahr2+1900.
exe.
formats Zeile Jahr (F4.0) Importe InProd Lager Konsum (F8.1) .
save outfile="C:\CP193.sav".
exe.
```

```

GRAPH
  /SCATTERPLOT (MATRIX) =Importe InProd Lager Konsum
  /MISSING=LISTWISE
  /TITLE= 'Matrix zur Veranschaulichung'
          'von Multikollinearität'
  /FOOTNOTE= 'Chatterjee & Price, 19952, 192'.

```

前面在简单线性模型的阐述中已经解释了 GRAPH 命令。

```

CORRELATIONS
  /VARIABLES=InProd Lager Konsum
  /PRINT=TWOTAIL NOSIG
  /MISSING=PAIRWISE .

```

这里不再进一步解释 CORRELATIONS 语句。

```

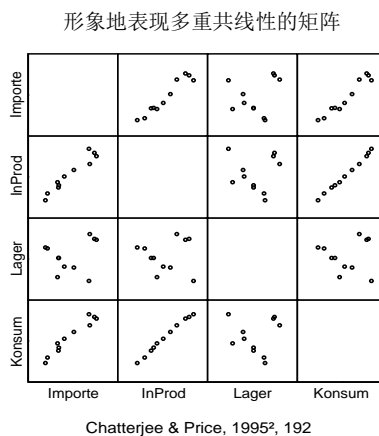
REGRESSION VARIABLES = InProd, Lager, Konsum, Importe
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS RANOVA COLLIN TOL ZPP
  /CRITERIA=PIN (.05) POUT (.10)
  /NOORIGIN
  /DEPENDENT Importe
  /METHOD=ENTER InProd Lager Konsum
  /PARTIALPLOT ALL
  /SCATTERPLOT= (*ZRESID, *ZPRED)
  /RESIDUALS HIST (ZRESID) NORM (ZRESID)
  /CASEWISE OUTLIERS (3) PLOT (ZRESID)
  /SAVE ZPRED COOK LEVER ZRESID DFBETA DFFIT .

```

前面在多元线性回归的阐述中已经解释了 REGRESSION 语句。

输出结果：多重共线性的迹象

多重共线性的第一个迹象是预测变量之间的线性关联，如 INPROD 和 KONSUM 之间。



这个矩阵给出了利用皮尔逊相关进行的推断性统计检验的相关统计量。

相关性				
		InProd (国内生产总值)	Lager (库存)	Konsum (消费额)
InProd	皮尔逊相关	1	0.026	0.997**
	显著性 (2 位数)		0.940	0.000
	N	11	11	11
Lager	皮尔逊相关	0.026	1	0.036
	显著性 (2 位数)	0.940		0.917
	N	11	11	11
Konsum	皮尔逊相关	0.997**	0.036	1
	显著性 (2 位数)	0.000	0.917	
	N	11	11	11

**相关性在 0.01 (2 位数) 的水平上是显著的。

预测变量 INPROD 和 KONSUM 之间的关联是近乎完美的线性正相关 (0.997, P = 0.000; 最后还额外计算了从 INPROD 到 KONSUM 的回归, 见下文)。多重共线性的成因可能比只有两个变量的简单线性组合的成因更为复杂, 当有多个变量时, 这个影响可能无法通过散点图矩阵容易地识别出来 (参见概述一节的说明)。

模型	R	R ²	调整 R ²	估计值的标准误差
1	0.996 ^a	0.992	0.988	0.4889

a. 预测变量: (常量), 消费、库存和国内生产总值。

b. 因变量: 进口额。

调整 R²表明, 解释方差的分量几乎完全得到了解释 (99.2%)。因为只有几个变量, 所以这个结果就更为重要。然而, 少数个案 (n = 11) 使测定的调整 R²具有相对性。

系数 ^a											
模型		非标准化系数		标准化系数	T	显著性	相关性			共线性统计量	
		B	标准误差	B			零阶	偏相关	部分相关	允差	VIF
1	(常量)	-10.128	1.212		-8.355	0.000					
	InProd	-0.051	0.070	-0.339	-0.731	0.488	0.965	-0.266	-0.025	0.005	185.997
	Lager	0.587	0.095	0.213	6.203	0.000	0.251	0.920	.211	0.981	1.019
	Konsum	0.287	0.102	1.303	2.807	0.026	0.972	0.728	.095	0.005	186.110

a. 因变量: 进口额。

表“系数”分别含有共线性的一个明确迹象和一个隐含迹象。VIF 值明显高于极限值 10, 因此非常清楚地表明了存在共线性。INPROD 前面的负号是一个隐含的迹象, 这个出人意料的负号与原先预期的正号相反。多重共线性可以清楚地解释这种现象。利用非标准化系数可以得出如下回归方程:

IMPORTE = - 10.128 - 0.051*INPROD + 0.587*LAGER + 0.287*KONSUM.

因此，由多重共线性而产生的、与预期相反的非显著性影响也输入了方程。人们的预期是，这个方程只是由于多重共线性而在统计学上显得有人为痕迹，但除此之外，根据经验还是可信的。因此，只要在使用（解释）这个方程时考虑到多重共线性的影响，就能得出实用的预测（参见下一节）。

共线性诊断^a

模型	维度	特征值	条件指数	方差分量			
				(常量)	国内生产总值	库存	消费额
1	1	3.838	1.000	0.00	0.00	0.01	0.00
	2	0.148	5.086	0.01	0.00	0.94	0.00
	3	0.013	17.073	0.77	0.00	0.03	0.00
	4	5.447E-5	265.461	0.22	1.00	0.02	1.00

a. 因变量：进口额。

表“共线性诊断”进一步说明了多重共线性的类型。例如，维度 4 中的特征值小于 0.01，因此清楚地表明具有共线性，相应的条件指数同样也是共线性的迹象。“方差分量”是指某个估计值的方差分量，这个估计值是通过与特征值相应的维度测定得出的。如果一个条件数量较高（例如，维度 4）的维度对两个或多个变量的方差构成巨大影响，则共线性就成了值得注意的问题（INPROD，KONSUM）。

处理多重共线性：三种方法

本节介绍了处理多重共线性的三种方法。这里主要采用了 Chatterjee & Price（1995²，195-196，210-212）著作中的表述。

从回归方程中剔除（假设）

如果只有两个相互关联几乎完美的变量，可能就很容易消除多重共线性。

由于两个相互关联的变量（INPROD，KONSUM）都不与模型中剩余的第三个变量（LAGER）有关联，因此可以分别将这两个变量中的一个从模型中剔除，然后分别计算两次到 IMPORTE 的回归分析（INPROD，LAGER、KONSUM，LAGER），而无需担心出现多重共线性。这种方法无须再多做解释。采用这种方法时的假设是，两个分别只带有两个预测变量的模型足以实现分析目的。

但是，如果在回归方程中仍然留有共线性的预测变量，则要使用另外两种方法。

保留在回归方程中（假设和加权）

例如，当有两个相互高度相关的变量时，利用一个可信的因果模型，更加准确地观察这两个变量之间的关系，可以得出很大的信息量。经过初步思考，国内生产总值（INPROD，预测变量）应对消费（KONSUM，因变量）规模的大小有影响。而相反方向的影响则几乎看不出来。

```
REGRESSION VARIABLES = InProd Konsum
/MISSING PAIRWISE
/STATISTICS COEFF OUTS R ANOVA COLLIN TOL ZPP
```

```
/CRITERIA=PIN (.05) POUT (.10)
/NOORIGIN
/DEPENDENT Konsum
/METHOD=ENTER InProd
/PARTIALPLOT ALL
/SCATTERPLOT= (*ZRESID, *ZPRED)
/RESIDUALS HIST (ZRESID) NORM (ZRESID)
/CASEWISE OUTLIERS (3) PLOT (ZRESID)
/SAVE ZPRED COOK LEVER ZRESID DFBETA DFFIT .
```

系数 ^a											
模型		非标准化系数		标准化系数	T	显著性	相关性			共线性统计量	
		B	标准误差	B			零阶	偏相关	偏相关	允差	VIF
1	(常量)	6.259	3.335		1.876	0.093					
	InProd	0.686	0.017	0.997	40.448	0.000	0.997	0.997	0.997	1.000	1.000

a. 因变量：进口额。

该回归方程是： $KONSUM = 6.259 + 0.686 * INPROD$ 。据此，国内生产总值所占消费的比例约为 69%。如果有充分理由假设两个相互关联的预测变量之间的比例是常量 $KONSUM = 0.69 * INPROD$ （例如，今后也保持恒定，这一点对于预测尤为重要），则这些信息也可以用来解释完整的回归方程。

下面对于完整的回归方程做出了两个解释。第一个解释忽略了多重共线性的影响，第二个解释则考虑到了这种影响。这些例子表明，在回归方程中可能留有一个非显著的影响及其成因，以及忽视多重共线性会导致对回归方程的错误解释。为了可靠地解释回归方程，准确地了解需要建模的对象是必不可少的。

完整的回归方程（见第 2.3.3 节“输出结果：多重共线性的迹象”）是：

$$IMPORTE = -10.128 - 0.051 * INPROD + 0.587 * LAGER + 0.287 * KONSUM。$$

方法 1：不考虑多重共线性（错误解释）

为了做出预测，例如，我们可以假设， $INPROD$ 将来会增加 10 个单位，而所有其他预测因素保持不变（零）。这样就得出了下列回归方程：

$$\begin{aligned} IMPORTE_{现在} &= -0.51 * INPROD_{现在} \quad \text{和} \\ IMPORTE_{未来} &= IMPORTE_{现在} - 0.51。 \end{aligned}$$

这个两个回归方程含有一个明显的错误和一个隐藏的错误。明显的错误是国内生产总值和进口额之间的负相关。从经济学角度来看，国内生产总值增加会导致进口额下降（-0.51）的这种预测是完全与实际相悖的。隐藏的错误可能就是这种现象的原因所在，也就是说，向这个方程代入数值时没有考虑到 $KONSUM$ 和 $INPROD$ 之间 1 : 0.69 的比例，而是将这两个数值之间偏离数据结构的，也就是人为造成的比例误以为真。

方法 2：考虑到多重共线性（正确的解释）

为了做出预测，例如，我们可以假设，INPROD 上升 10 个单位，从而由于多重共线性的存在使 KONSUM 增加大约 14.5 个单位（ $14.5 = 1/0.69 * 10$ ；也可参见下面的备注）。LAGER 保持不变（零）。这样一来，代入的数值考虑到了变量之间的多重共线性关系，从而得出了下列回归方程：

$$\text{IMPORTE}_{\text{现在}} = -0.51 * \text{INPROD}_{\text{现在}} + 4.16 * \text{KONSUM}_{\text{现在}}$$

$$\text{IMPORTE}_{\text{未来}} = \text{IMPORTE}_{\text{现在}} - 0.51 + 4.16$$

$$\text{IMPORTE}_{\text{未来}} = \text{IMPORTE}_{\text{现在}} + 3.65$$

备注：数值 14.5 是通过 1 除以 0.69（假设的恒定比例）再乘以假设 INPROD 上升的 10 个单位得出的。数值 4.16（ $4.16 * \text{KONSUM}$ ）是通过 0.287 乘以之前测定的 14.5（见上面的初始方程）得出的。

国内生产总值上升以及可以考虑到的消费增加，共同导致进口额上升（3.65）。

从方法 1 和方法 2 得出的两个完全不同的结果可以看出，在解释多重共线性数据的回归方程时需要特别谨慎。

方法 3：将多重共线性预测变量加权

在上一节中我们知道，根据经验 KONSUM 和 INPROD 之间的比例为 0.69。如果要替代两个共线性预测变量中的一个，则可以将这个加权代入初始方程。根据假设的因果方向，在本例中替代了 KONSUM。

初始方程（具有共线性）：

$$\text{IMPORTE} = -10.128 - 0.051 * \text{INPROD} + 0.587 * \text{LAGER} + 0.287 * \text{KONSUM}。$$

新方程（无共线性）：

$$\text{IMPORTE} = -10.128 - (0.051 + (0.69 * 0.287)) * \text{INPROD} + 0.587 * \text{LAGER} \text{ 和}$$

$$\text{IMPORTE} = -10.128 - 0.249 * \text{INPROD} + 0.587 * \text{LAGER}。$$

新的方程中不再含有多重共线性。通过这种方式测定的 INPROD 影响权重不仅包含自身的影响，而且还包含 KONSUM 的影响。如上所述，这个模型所基于的假设是，相互关联预测变量之间的比例恒定保持在 1 : 0.69。

另一种方法是分量设计，如通过主分量分析或偏回归，然后，就可以替代相互关联的自变量而将测定出的正交分量代入模型。第 5.1 节详细介绍了各种类型偏回归法（PLS）。

2.4 计算线性回归的前提条件

计算线性参数回归的前提条件是十分宽泛而复杂的，这方面可以参见 Cohen 等人（2003³）、Chatterjee & Price（1995²）和 Pedhazur（1982²）的著作。

在计算回归时，应注意下文汇总的各个前提条件。其中主要包括因变量的定距型数据，以及线性相关和正态分布的残差。预测变量的彼此独立性是多元回归的一个特殊的前提条件，也

就是说,解释变量相互之间不允许具有多重共线性(主要是通过特征值、VIF 和条件数可以识别出来)。对于非独立数据,不允许存在残差的自相关(参见杜宾-瓦森)。对于使用时间相关数据并且含有多个自变量的模型,应一次性地对多个前提条件进行检验(例如,自相关和多重共线性)。

对于各种变量选择方法(如方差膨胀因子)的应用和解释、各种不同方程估计值(如 C_p)是否等同的重要评估标准的应用和解释,或者究竟是遵守还是破坏了模型假设的讨论,通常可以参见 Cohen 等人(2003³, 117-150)、Chatterjee & Price (1995², 265-267)和 Pedhazur (1982², 221-254)的著作。

1. 回归分析的目的是确定无疑的:描述、估计、预测和模拟等不同的目的对变量数量和模型拟合优度(主要是预测性能与模型准确性的对比)提出了各种不同的要求。因此,应用目的就确定了之后在测定回归方程时可以优化哪些方面(Chatterjee & Price, 1995², 245-246):在(a)描述时,应该只利用最小数量的变量就可以解释绝大部分的因变量方差;在(b)预测和诊断时,应利用最小均方预测误差来选择变量;对于(c)模拟而言,应该尽可能准确地估计预测方程中的变量的回归系数,因此其标准误差应该尽可能小。在有些情况下,一个回归方程也要同时满足多个要求。

当然,只有当数据的测量是完美无缺的,并且完全满足了回归分析的所有前提条件时,这种微调才有意义。

2. 回归分析假设了(至少)一个自变量(X)和因变量(Y)之间的因果模型。根据逻辑,从一开始就剔除了伪回归。例如,送子仙鹤数量对新生儿数量的影响。在模型中只给出了有关变量,无关变量也应予以剔除。
3. 变量都是定距的。既可以根据项目文档、数据视图、变量视图(只要给定的测量水平是正确的)和探索性描述分析(如频率分析),也可以通过图形分析(如散点图)确定是否有定比变量的测量值序列。原则上,也可以将 1/0 编码的哑变量(包括其交互作用)引入回归分析,这不仅可以使回归分析与(M) AN (C) OVA 等效,在某些情况下(例如,不相等的方格频数),甚至可能优于(M) AN (C) OVA (Pedhazur, 1982², 271-333; Chatterjee & Price, 1995², 99-125)。但是,引入分类变量会带来多重共线性的风险(参见 Pedhazur, 1982², 235)。
4. 必须针对同一个对象成对地测量 x_i 和 y_i 。换言之,所调查的特征是取自一个样本的相同元素。个案必须是相互独立的。回归分析不适合用于对相关数据进行分析(参见自相关,见下文)。对于观察数据,应确保尽管这些观察数据具有地理定义或其他取决于形势的定义,但仍然可以认为对某个总体及其结果具有代表性。
5. 样本量:对于确定所需的样本量 N ,存在着不同的标准。个案(观察)的数量应该超过自变量或需预测参数的数量。基本的经验法则是: $N \geq 50 + 8m$ (m 表示模型中所期望的预测变量数量)。预测优度越好,模型性能越强大;或者功效越小,则需要 的个案越多。

例如,Green (1991)就提出了一种估计方程,其中同时考虑到了所期望的功效: $N \geq (8/f^2) + (m-1)$ 。 m 表示模型中所期望的预测变量数量。根据 Cohen 等人著作(2003³), f^2 表示所期望的功效,例如,0.01、0.15 和 0.35 分别代表效应小、中等、大。例如,如果一个模型在功效中等的情况下含有 5 个预测变量,则通过方程 $N \geq$

$(8/f^2) + (m-1)$ 可以计算出, 至少需要 $N = 58$ 个个案。

也有个案数量太多的风险。如果个案过多, 则 R 和 R^2 总是显著地区别于零, 并预测因变量具有非常微不足道的变化。这无论是从统计学还是实践角度来看都是非常不利的。如果变量数量超过个案数量, 则也可以使用偏回归 (参见第 5.1 节)。

6. 在因变量和自变量的测量值之间存在着已知的关联。在线性模型中, 这种关联是线性的。在非线性模型中, 则函数必须是已知的 (例如, 余弦、对数、指数等)。在这个函数的背景下, 两个变量是高度相关的。例如, 根据皮尔逊相关系数, 两个变量的关联是线性或者线性化的。
7. 离群值: 离群值不存在或者存在时应予以删除。回归分析对离群值的反应十分敏感。即使很少的离群值 (残差或者高杠杆值) 也足以对回归分析结果 (对回归权重的估计精度) 产生深远的影响。例如, 在分析之前, 可以通过系统的描述性分析将明显的 (单变量) 离群值识别出来 (规则: 离群值是距离平均值超过 4.5 个标准差的数值), 在分析之后, 可以通过很高的马氏距离识别出 (多变量) 离群值。在删除或者转换离群值之前, 应小心地对其进行检验。形式上的显著数值并不一定是内容上的显著数值 (尤其是在社会科学领域的数据中), 而且也不能排除这种可能性: 检验统计量本身就是不可靠的。
8. 缺失数据 (缺失值): 在制定预测模型时, 缺失数据可能会导致问题。预测模型的理想条件是不缺失任何数据。如果数据是完全随机缺失的, 则具体的缺失程度决定了分析时还留有多少百分比的数据, 这可能还会导致出现问题。如果通过合理的考虑, 发现缺失值以某种方式与目标变量相关, 那么只要从模型中剔除了这些缺失值, 模型的解释和建模就会产生问题。例如, (a) 从建模角度通过一个指示缺失值的指标和 (b) 替换缺失数值, 从而将缺失数据重新引入模型。但是只能在这个前提条件下: 这些缺失数据的编码、重建和模型集成是合理的、可追溯的。如果缺失值集中在一个变量上, 则或许也可以从分析中剔除这些缺失值。
9. 测量误差: 应十分可靠地, 也就是在理想情况下没有误差地测得两个自变量。对于简单回归而言, 自变量中的测量误差会导致对回归系数的估计偏低。因变量中的测量误差不会影响对回归系数的估计, 但是会加大标准估计误差, 从而影响显著性检验。对于多元回归而言, 测量误差通常会导致对 R^2 的估计偏低。因变量中的测量误差可能会导致对回归系数的估计偏低, 而自变量中的测量误差则情况比较复杂, 会导致对回归系数的估计偏低或者偏高 (Pedhazur, 1982², 230-232)。自变量最好是正态分布的, 但这不是必需条件。
10. 残差。对于误差有下列假设: 残差的平均值等于零。残差的方差是恒定的 (方差齐性, 同方差性, 参见下文)。若误差和因变量的预测值之间的关联是随机的、误差与自变量不相关 (参见偏回归图)、应用在图形式残差分析上, 则这些意味着: 在样本足够大的前提下, 一个模型的预测误差 (残差) 是线性和正态分布的, 并且围绕在平均值 0 周围散布 (正态性假设) 且不存在异方差性。在偏回归图中, 误差与自变量不相关 (参见 Pedhazur, 1982², 36-39; Chatterjee & Price, 1995², passim)。
11. 同方差性假设: 尤其对于多元回归而言, 方差一致性 (又称同方差性) 假设是其核心的假设。对于因变量的所有预测值, 误差的方差是恒定的。在残差图中, 可以从这个

现象识别出异方差性：散布面不是矩形对称的，而是从左向右呈剪刀状张开的（前提是样本足够大）。

12. 测量值变异：特征量。例如， R^2 、F 值和回归方程的显著性，以及如多重共线性等现象都主要受数据的值域和数值变异的影响（参见 Chatterjee & Price, 1995², 190; Pedhazur, 1982², 30-32）。额外的或者经过补偿的数据可能对这些指标有决定性影响。回归函数应既没有超出现有的测量值域，也不应在测量值域之外被解释。
13. 从概念和计算的角度应删除特殊的回归效应。例如，在双变量回归中应确保关联只是由两个受调查的变量，而不是由其他变量造成的（如偏相关/偏回归）。在有多个预测变量时，主要是消除多重共线性，确定抑制变量的效应。对于双变量和多变量的模型变体，则应消除自相关。

如果预测变量由于与其他预测变量的相关而抑制了其用于预测因变量的无关方差，从而增大了其回归系数或者 R^2 ，则这些预测变量就称为抑制变量。如果有多个变量，则从中识别出抑制变量并不简单，主要是因为抑制效应有多种不同的类型（参见 Cohen 等人的著作，2003³）。第一个依据是预测变量明显的相关系数和回归系数。如果与因变量的相关性在数值和正负号大致等于一个预测变量的回归系数，则表明这个预测变量可能是抑制变量。如果从回归方程中删除这个变量，并且其他变量的回归系数明显变差，则就可以成功地将被删除的变量识别为抑制变量。

14. 尤其是在解释关于心理测量尺度的回归结果时，应注意变量的极端值。
15. 对于某些提出的问题应注意，回归直线究竟是可以穿过 Y 轴的任意一段，还是应从零点出发。在对方程中带有和不带截距的模型进行比较时应注意， R^2 统计量由于具有不同的成立方式，因此不适合用于对模型的比较。
16. 设定误差（建模）：尤其是在利用多元回归检验多变量模型时，可能出现设定误差（参见 Pedhazur, 1982², 225-230, 251-254）。如果一个形式上显著的模型无法保持在关于内容的理论所设定的位置上，就存在设定误差。通常的错误包括：剔除了有关变量，进入了无关变量或者将曲线关联设定成线性关联。
17. 在建模时，理论起到关键作用。根据制定的模型不同，预测变量可以证明是有效的或者是无效的。不同的理论决定了不同的模型，而这些模型又可以对各个预测变量做出不同的证明。涉及预测变量的现象表明，一个预测变量的影响不仅取决于模型拟合优度（测量误差）和所基于样本的特征，而且还取决于每次制定的或者检验的模型，因此可以视为是相对的。
18. 建模应是通过关于内容的统计学标准，而不是通过关于形式的算法来推导的。很多作者明确建议不要使用自动变量选择的方法，但是在有保留的情况下，作为一种探索性方法也是可以使用的。对于这两种不同的工作方法，建议采用下面的处理方式：首先进入一个内容上有关的预测变量，然后通过显著性检验（T 统计量）剔除统计学上无关的变量。
19. 逐步法：逐步法是建立在关于形式的标准基础上的，因此不适合用于理论推导的建模。应根据可信的、关于内容的标准或者交叉验证，对纯粹探索性的或者预测性的工作方法进行交互检验。与前进法相反，在部分考虑到可能的交互作用的情况下，后退法可

以识别出一系列方差解释程度最大的变量 (Mantel, 1970)。但是, 逐步法不消除多重共线性, 因此至少要通过交叉验证予以保障。

20. 多重共线性: 多重共线性是多元回归的一个特殊问题, 主要可以通过皮尔逊相关 (例如 >0.70)、VIF (允差)、特征值和条件数识别出来, 但是也可以通过与期望相反的正负号或系数的标准误差识别出来 (Chatterjee & Price, 1995², 184-196、197-203, 258-260)。如果模型中预测变量的所有 VIF 都低于 10, 则共线性就没有问题。同样, 当条件数较小 (<15) 和特殊值较大 (>0.01) 时也是如此。但是当条件数 >30 时, 就应采取措施。

应首先识别出一个令人满意的模型, 并在删除了残差或者高杠杆值后, 才采取消除多重共线性的措施。出现多重共线性第一个原因可能是设定误差: 应识别出造成多重共线性的变量, 并尽量从模型中予以删除。一个经常造成多重共线性的原因是同时引入了多个等效的指标 (例如, 多次智力检验或者类似的计量经济学指标), 而且哑变量的特定编码也可能造成多重共线性。出现多重共线性的第二个原因可能是样本错误: 只抽取了很少, 但是相互相关的数据组合。解决的办法是再采集其他的数据组合。多重共线性的第三个原因可能是变量之间的关联具有一种所调查对象固有的特性。这个现象无法通过额外的数据予以补偿。在这种情况下, 可以通过偏回归或者岭回归查找能够解释预测变量之间关系的因子 (Chatterjee & Price, 1995², 203-207, 221-228, 230-235)。

21. 自相关: 回归的另一个假设是残差的独立性。相反, 残差的相互关联就称为自相关, 表示前后连续个案的残差 (大多数情况下是在时间上的, 但通常也是在空间上的) 的关联性。

自相关在大多数情况下取决于调查的设计。如果是同时获取数据的横向调查研究, 则自相关是无关的。但如果是先后依次获取数据的纵向调查研究, 则必须检验自相关的规模 (参见 Chatterjee & Price, 1995², 179-181)。例如, 在纵向数据中, 自相关表现为隐藏的, 经常是时间相依的或者季节性的结构或者影响; 在横向数据中, 可能有的自相关是数据组织的人为臆造, 大多数情况下是无关的。

自相关是一个值得重视的问题。除了使残差产生偏误之外, 自相关的后果主要还有无效的显著性检验、无效的置信区间和不准确的估计。自相关的原因可能是数据问题 (例如, 时间依存性或者其他因果依存性, 也就是人们所说的“真实”自相关) 或者设定误差 (非线性代替线性关系, 方程中缺少有关变量, 也就是人们所说的“虚假”自相关)。

杜宾-瓦森统计量用于检验一阶自相关, 也就是说检验一个残数是否依存于其先前的一个残差。杜宾-瓦森统计量对零假设 (自相关等于零) 和备择假设 (自相关不等于零) 进行检验。杜宾-瓦森统计量的值在 0~4 之间。检验值 d 越靠近 2, 则越能说明不存在自相关。检验值与 2 偏离则表明存在自相关。检验值 >2 表明存在负自相关 (正值伴随负残差, 或者负值伴随正残差), 检验值 <2 则表明存在正自相关。SPSS 目前没有输出对杜宾-瓦森统计量的显著性检验。为了判断某些带有极限值 L_U 、 L_O 或 $4-L_U$ 、 $4-L_O$ 的无差异区域是否超过了显著性水平, 必须查看杜宾-瓦森表。为此, SPSS 在一个单独的文档中提供了根据 Savin 和 White 的方法 (带有截距的模型) 或者

Farebrother 的方法（没有截距的模型）对 $N = 6$ 至 $N = 200$ 个样本进行检验的杜宾—瓦森临界值。通过适当地转换相关数据，可以清除“真实”自相关。通过检验和校正可能有的设定误差，可以清除“虚假”自相关（Chatterjee & Price, 1995², 163-168）。

如果有自回归，则可以变换成偏回归或者带有自回归误差的回归分析（例如，利用 Yule-Walker 估计值）。对于具有时间相依的或者季节性的结构或影响的纵向数据，可以使用 OLS 回归（例如，参见 Woolridge 的著作，2003，第 10 和 11 章、Cohen 等人的著作，2003³，第 15 章、Chatterjee 和 Price 的著作，1995²，第 7 章），也可以使用时间序列分析的特殊方法（例如，参见 Hartung, 1999，第 XII 章；Schlittgen, 2001；Schlittgen & Streitberg, 2001⁹；Yaffee & McGee, 2000）。

22. 交叉验证作为预测优度的检验方法（排除过度拟合）：一个模型在完成参数化之后，应利用交叉验证检验其预测优度是否具有实际的关联性。如果利用在建立模型时所应用的样本（称为“训练数据”或者“训练样本”）来检验这个模型，则可能就将命中率估计过高（过度拟合）。在十分特殊的模型中，经常会出现过度拟合现象。原因通常是训练数据集（乖离率、分布等）的特殊性。因此，应根据经验数据（如 20% 的数据），始终通过交叉验证检验模型中是否存在过度拟合。交叉验证是利用一个或者多个其他（子）样本（称为“验证数据”）对模型进行的检验，在 REGRESSION 中可以很方便地通过 /SELECT 选项调用。通过下列语句，可以将这些数据事先分解为用于训练和验证的子数据集。

```
compute TRAINING= (uniform(1) <=.80) .
variable label TRAINING 'Trainingsdaten (ca. 80%)' .
exe.

value label TRAINING
1 'Trainingsdaten'
0 'Validierungsdaten' .
exe.
```

注意：在 UNIFORM 函数中应注意，每次实施计算时可能都会得出不一样的结果，有时可能距离设定值（如 80%）比较近，有时可能比较远。

拟合不足则意味着模型中没有包含所有有关变量（措施：必要时将其引入模型）。过度拟合则意味着模型中也包含了无关变量（措施：必要时将其剔除）。

如果一个模型表现出很大的性能差异，例如，用训练数据可以将 80% 的数据正确分级，但是利用经验数据时这个比例可能只能达到 50%，则就存在过度拟合。相反的现象就称为拟合不足，也就是忽视了真实的数据现象。拟合不足现象主要发生在过于简单的模型中。

第 3 章 逻辑回归和有序回归

第 3 章介绍了逻辑回归和有序回归的基本方法。本章的结构是根据因变量的尺度水平构建的。最后几节分别归纳了所介绍方法的各种前提条件，以及对其进行检验的方法。

第 3.1 节首先通过一个总览表归纳了选择 SPSS 中所包含的方法，这些方法可能适用于对带有定类因变量的模型进行分析（参见第 3.5 节，在这里也要参照第 6 章中的说明）。

二元逻辑回归（SPSS 过程命令 LOGISTIC REGRESSION，第 3.2 节）需要使用一个分为两级的因变量，这个方法中没有考虑因变量中的极差信息。第 3.2 节首先介绍了作为基本方法的二元逻辑回归，然后阐述了这种方法与其他方法的共同点和区别（主要是模型、尺度水平），并根据几个计算实例，主要阐述了变量选择的不同方法，以及对所输出统计量的解释。最后探讨了经常出现的模型拟合优度和预测精度不一致问题。

有序回归（SPSS 过程命令 PLUM，第 3.3 节）需要使用一个最少二级的（定序）因变量，并且考虑到了因变量中的极差信息。然后阐述了这种方法与其他方法的共同点和区别（主要是模型、尺度水平），并根据几个计算实例，主要阐述了如何解释模型的 SPSS 输出结果，其中这些模型带有定距和分类预测变量。

多项逻辑回归（SPSS 过程命令 NOMREG，第 3.4 节）同样需要使用一个至少二级的定类因变量，这种方法没有考虑因变量中的极差信息。对多项逻辑回归的阐述与第 3.2 节类似。此外，还介绍了一种特殊情况，即带有定量预测变量的巢式病例对照研究（1：1）。

3.1 引言：因变量的因果模型和测量水平

分类回归（又称定性回归或者离散回归）模型的应用目的和定量回归模型一样，即尽量简单、参数需求量尽量小地表现一个因变量和一个或多个自变量之间的关联。应可以评估各个预测变量的关联性。对于预测变量的特定组合，应可以尽量好地预测因变量。尤其针对分类回归的模型评价（主要是残差分析和偏差分析），可以参见 Hosmer & Lemeshow（2000）、Menard（2001）和 Tutz（2000）的著作。

形式最简单的分类回归调查的是两个变量之间的关联。与（非参数）参数相关或者（对称）相关不同的是，回归的前提条件是一个因果模型，即（至少）两个变量之间的因果关系。回归的目的是，调查一个变量与一个或多个自变量之间是否存在因果依存性。

此时因变量被称为目标变量、回归应变数或标准，预测变量通常也称为预测量、自变量或者回归量。如果模型中只含有一个自变量，则被将其称为简单回归；如果模型中含有多个变量，则将其称为多元回归。

SPSS 中实现的建模每次只允许用一个因变量，并且只能在一个层次上进行分析。必要时可以利用路径分析计算带有多个因变量，并且在多个层次上的回归。同样是 SPSS 公司出品的 AMOS 系统可以对这样的或者类似的问题进行计算，并将多个层次和多个因变量引入一个共同的分析模型。

分类回归的类型主要是通过模型实现区别，也就是通过因变量的尺度水平和分类回归所依据的数学模型，具体而言这种区别主要体现在自变量的尺度水平上（详细说明参见本章结尾）。

下面介绍了几种类型的分类回归，尤其是对现有数据进行定量回归的一个或多个前提条件没有得到满足时，可以选择这几种分类回归。为了避免产生误解，下面介绍的分类回归方法从广义上来讲属于参数方法（参见 Tutz, 2000、Böhning, 1998 和 Chatterjee & Price, 1995²）。

属于分类回归的，并且在 SPSS V16 的“回归”模块中提供的，主要有二元逻辑回归、多项逻辑回归、有序回归和除了概率单位回归之外的对数线性模型。不能把上述“分类回归”概念与特别的 SPSS 过程命令 CATREG（通过“Categories”模块获得许可，通过“回归”→“最优尺度”调用）混淆。CATREG 执行的是狭义上的分类回归。CATREG 使用的是基于交替最小二乘法的最优尺度分析。如果 CATREG 向定类数据分配数值，则可以将定类数据量化、标度，并像定量变量一样引入一个线性回归模型。由于具有“最优尺度”过程，因此 CATREG 与其他真正的回归分析方法有根本性的区别，这里不再赘述。

介绍的定类数据回归方法

方法	自变量	自变量尺度
二元逻辑回归	系数：分类 协变量：定距	两个类别（二元）
多项逻辑回归	系数：分类 协变量：定距	超过两个类别（分类变量）

续表

方法	自变量	自变量尺度
有序回归	系数：分类 协变量：定距	超过两个类别（定序变量和离散变量）
概率单位回归	系数：分类 协变量：定距	二元 特点：数值不得为负
对数线性分析，例如 Poisson 对数，Logit 对数线性分析	系数：分类 协变量：定距	分类变量
偏回归（PLS）	系数：分类 协变量：定距	类变量，定距变量
分类回归	有序样条，标称样条，分类变量，定序变量，定量变量	有序样条，标称样条，分类变量，定序变量，定量变量

下面，首先详细阐述和解释对二元逻辑回归的计算，然后是对两种有序回归的计算。最后介绍多项逻辑回归的方法。

由于 SPSS 过程命令 LOGISTIC REGRESSION 和 NOMREG 的几个特点，建议所有对二元逻辑回归方法感兴趣，以及所有原本只对多元回归方法感兴趣，但也想了解二元逻辑回归的读者参见多元回归的一章。在本章结尾简单地介绍了其他回归类型（也可参见第 6 章）。

在关于 SPSS 过程命令 NOMREG 的一节，也介绍了对巢式病例对照研究的分析，这种形式的分析无法用 SPSS 过程命令 LOGISTIC REGRESSION 实施。

3.2 二元逻辑回归

二元逻辑回归（又称逻辑判别）的特点主要是在因变量中体现事件发生与否的二元（“非此即彼”）性，举例如下。

- 根据哪些实验室参数，可以将患者归为症状 A 或者症状 B？
- 哪些产品特点会使产品销售情况良好或者不好？
- 是否可以根据预测变量，可靠地将人识别为“已婚”或者“单身”？
- 在个人出行时，是根据哪些因素选择乘坐近程公共交通工具或者自己开车？
- 根据资产负债表上的哪些参数，可以将一个银行客户归类为信贷资质良好客户或者不良客户？

这些问题都是二元逻辑回归的典型应用实例。在下文中，逻辑回归通常指的就是定距预测变量的二元逻辑回归和二元因变量的二元逻辑回归。在备注中，阐述了离散尺度预测变量在建模、编程和解释方面的特点。

3.2.1 逻辑回归方法和与其他方法的比较

与分类回归的其他方法相比，逻辑回归具有很多区别或者共同点（参见本章末尾的一览表）。

二元逻辑回归与有序回归的区别在于，二元逻辑回归没有考虑到因变量中的极差信息，并且因变量只能假设有两个可能取值。二元逻辑回归的扩展形式就是多项逻辑回归，可以选择考虑或者忽视因变量中的极差信息，并且这个因变量具有两个或者多个可能取值。如果绝大部分是定量预测变量或者完全是定量预测变量，则优先使用二元逻辑回归，相反，如果绝大部分是分类预测变量或者完全是分类预测变量，则优先使用多项逻辑回归。因此，对于 0/1 编码的因变量和离散尺度的预测变量，LOGISTIC REGRESSION 和 NOMREG 会得出相同的结果（例如，Wald 统计量），只是在过程特定的 SPSS 输出结果上有所不同。相反，对于 0/1 编码的因变量和定比预测变量，LOGISTIC REGRESSION 和 NOMREG 会得出不同的结果，因为 NOMREG 是建立在单个个案的基础上，而 LOGISTIC REGRESSION 使用的是成组数据。因此，LOGISTIC REGRESSION 应优先用于定比自变量，而 NOMREG 主要用于绝大部分是或者完全是离散尺度的预测变量。

当逻辑回归本身的前提条件得到满足，而判别分析的特殊前提条件无法全部或者甚至完全不能得到满足时，逻辑回归可以视为判别分析的一种有力的替代方法。例如，在组群之间差异特别大、没有给定多变量正态分布或者二元预测变量的情况下，就应优先使用逻辑回归，而不是判别分析（例如，参见 Klecka, 1980, Press & Wilson, 1978）。只要满足了所有前提条件，就应优先使用判别分析而不是逻辑回归。但是在这里有一个限制：当有二元预测变量时，判别分析可能会有对关联估计过高的倾向（Hosmer & Lemeshow, 2000, 22, 43f.）。新一些的逻辑回归替代形式如分类树（如 SPSS' AnswerTree）、神经网络（如 SPSS' Clementine）或者非参数回归（广义加性模型，GAM）。

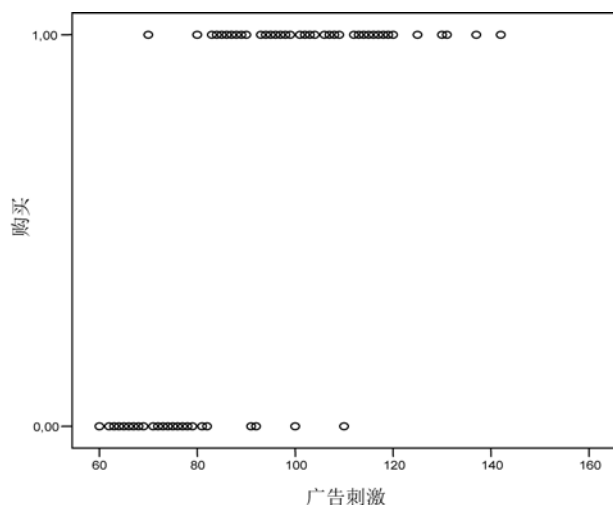
根据 Rothman 和 Greenland (1998) 的著作，逻辑回归模型被视为流行病学最常用的模型之一，与其他类型的分类回归共同对一个独立的科研领域，即生物测定学的发展做出了贡献。这个学科主要是调查 S 形剂量-效应关系、浓度-效应关系和时间-效应关系。在本章结尾，介绍了逻辑回归，以及各种类型方法的前提条件。还有几段从理论上和实践上比较了统计方法，或者 SPSS 过程命令“二元逻辑回归”（LOGISTIC REGRESSION）和“多项逻辑回归”（NOMREG）的区别和共同点。在菜单或者语句中，这两个 SPSS 过程命令的名称可能会让人糊涂：通常，只有当预测变量是定距（定量）时，才使用（二项、多项）逻辑回归这个名称；如果预测变量是分类预测变量，则统计学上正确的名称是“逻辑回归模型”（也可叫作“Logit 模型”）。

利用逻辑回归可以解决提出的下列问题：检验预测变量的关联性（必要时还要检验预测变量相互之间的交互作用或者预测变量与协变量的交互作用），估计预测变量，预测一个事件或者分类，以及一个分类模型的拟合优度。

通过基于一个二元因变量（“事件发生”或者“事件不发生”）和一个定量预测变量（如鼓励）的简单例子，我们可以想象出逻辑回归的逻辑。例如，超市的顾客在新品上市促销活动时总会受到某些广告宣传的影响，顾客购买相应商品就视为“事件发生”，不购买就视为“事件不发生”。将接受促销活动激励的程度记录为定量预测变量，例如，广告强度指数（广告规模，如灯光、声音）、持续时间指数（广告长度），或者是经济上的刺激（如价格）。

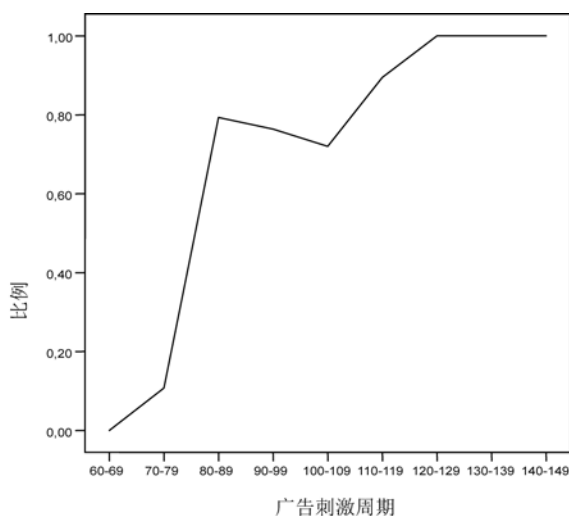
如果比较“购买”和“不购买”两组数据，则可以发现，“购买者”一组受到了更强的广告宣传刺激，因此在这一组中表现出较高的广告刺激值（“1”，上方直线），而在“不购买”（“0”，下方直线）一组中现出较低的广告刺激值。

在右侧的散点图中，两个独立的数据序列图明了两组广告刺激值的绝对值。



因此，在广告刺激周期较长时，“购买者”的比例较高，“不购买者”的比例相应较低。

现在，如果将广告刺激尺度转化为较为粗略的尺度（例如，以 10 为单位），并且测定在每个周期内购买者（“事件发生”）人数除以总人数（购买者和不购买者相加）的商，则可以在一个直线图中截取这个定量的相对频率（在不同周期内的商通常是基于不同大小的子样本）。右边的图形出现了两个组，其中一组的相对值，在这个个案中就是当广告刺激周期延长时增加的购买者。



上图是基于购买者与参加促销活动顾客总人数的比例，根据相互连接的数据点，反映了随着广告刺激周期的延长，购买者的比例逐渐增大。如果想在图中画出相反的比例，例如，不购买者与顾客总人数的比例，则可以看到一条递减曲线。

这种图通常反映了（曲线）线性函数，并且展示了描述二元变量和定量变量之间（曲线）线性关联的相对频率。逻辑回归建立在 0 和 1 之间的这个相对频率基础上，简而言之，将这个频率解释为在 0 和 1 之间的概率。在这个数值的基础上，可以根据自变量各自的水平（类别）得出相应类别因变量的发生概率。

逻辑回归与线性回归最主要的区别是，逻辑回归是建立在因变量尺度水平的基础上。线性回归的定距因变量可以让人对观察到的成对测量值进行分析，而在逻辑回归中，将发生一个目

标事件（“1”）的相对频率解释为每个级别自变量的条件概率。在一个或多个自变量的某个等级（等级组合）中，发生一个目标事件的频率越大，则相应变量等级（等级组合）产生影响的概率就越大。如果一个目标变量在某个变量等级（等级组合）中始终发生或者从不发生，则其概率等于 1 或者 0，否则这个概率值在 0 和 1 之间波动。

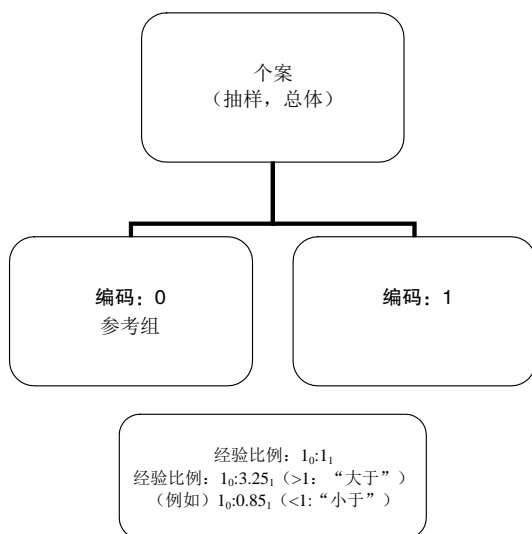
定量回归表现了由每次观察数值组成的成对测量值之间的线性关系，与此相比，逻辑回归则绘出了成对测量值之间的非线性（单调）关系，而这些成对测量值是由自变量观察值和测定的条件概率（“机率”）组成的。如果将这个概率对数化，就得到了与自变量呈线性相关的“对数”（是逻辑回归的主要前提条件之一），由此又能推导出很多与线性回归类似的特性（对此参见 Hosmer & Lemeshow, 2000、Menard, 2001、Tutz, 2000 和 Böhning, 1998）。逻辑回归也类似于协方差分析，因为这两种方法都是根据协变量（协方差）可能具有的效应进行调整的。协方差分析是基于定距数据的平均值，与此相比，逻辑回归是基于二元数据的比例。

示例

如果将实验室数值不同的人分为病人（“1”）和健康人（“0”），则也可以将这个区分表达为概率（“几率”）。例如，如果始终将实验室数值超过某个界限的人划分为病人，则可以说，当一个人的实验室数值超过这个界限时，将他归类为病人（“1”）的概率等于 1。从另一方面来说，如果将实验室数值始终低于某个特定界限的人划分为健康人，则将他归类为病人（“1”）的概率等于 0。在这两个界限之间的实验室数值的概率则在 0 到 1 之间波动。

通过将观察的测量值（自变量的水平）和条件概率（关于因变量的相对频率）构成一对，逻辑回归即可对一个或多个定距自变量和一个二元因变量之间的曲线关联进行描述和分析。

在这个背景下，可以根据预测变量的数值描述或者预测一个特性或者一个事件是否发生。逻辑回归的系数可以用于估计自变量的胜率。



上图表明，为什么在进行（二项）逻辑回归之前需要考虑在两组（有可能多组）之间进行比较的方向和编码。原则上，在比较时，例如，在两组（如 A 和 B）之间进行比较时，比较方向是开放的：A 可以与 B 比较，B 也可以与 A 比较。但是为了能够正确地计算输出的、表示比

例的参数估计值，则必须确定这两个组中的哪一个是比较的基础。根据一个比例是从 A 还是 B 的角度量化的，这个比例和相同的实证比例（理所当然）会得出另一个结果。例如，如果 A 相对于 B 的比值为 1:3，则从 B 的角度得出的、相同的实证比例就是 B 相对于 A 等于 0.333:1）。尽管实证比例是相同的，但是这个比值是从 A 还是 B 的角度来量化，（理所当然）会有很大区别。除了两个组的相互比值完全相等这种情况之外，根据 A 还是 B 构成比较的基础，逻辑回归始终会得出不同的胜率。为了可以正确解释两个组之间的比值，必须确定两个组中的哪一个是参考组（例如，病例对照研究中的“对照”），哪一个是对照组（例如，病例对照研究中的“病例”）。如何确定这个因素，会影响对胜率以及相应置信区间的估计。对于回归系数 B，主要是影响正负号，标准误差、自由度、Wald 值和显著性不受影响。

在下面这个例子中，用因变量的 0/1 编码计算了一个模型，用因变量的 1/0 编码计算了同样的模型。胜率和回归系数 B 发生了变化。

示例编码: 0/1					
回归系数 B	标准误差	Wald	df	Sig.	Exp(B)
1.449	0.344	17.742	1	0.000	4.259
示例编码: 1/0					
-1.449	0.344	17.742	1	0.000	0.235

但是两个组之间的比值没有发生变化。0 组相对于 1 组的比值为 1:4.259，1 组相对于 0 组的比值为 0.235:1）。胜率为各自的倒数，即 1:4.259 得出 0.235，1:0.235 得出 4.259。

例如，在临床或者流行病学研究中适用的规则是，始终将病例个案与对照个案进行比较（病例个案的因变量始终编码为“1”，对照个案的因变量始终编码为“0”）。因此，SPSS 的预设置就是始终用“0”（参考组）来比较“1”（对照组）。如果是另外的情况，则应在考虑到所提出问题（比较方向）的情况下，仔细检验因变量编码是否匹配相互比较的两个组。如果有其他编码，则在数据集中的数值层面上可以通过 RECODE 选项，或者在 LOGISTIC 语句层面上通过 FIRST/LAST 选项，使原本期望的比较方向匹配所提出的问题。

根据对所建立模型估计的胜率可以推导出，一个人（与各自的参考组相比）从一个组进入另一个组的概率有多大。

通常是用 0（参考组）来比较 1。

除了前面介绍的 1/0 编码（也称为 reference cell 编码）外，主要还有“平均值偏差”编码（如 1/-1）。编码方式影响了所测定的胜率以及置信区间的端点。但是 reference cell 编码一般认为比较容易解释（Hosmer & Lemeshow, 2000, 50-54）。

3.2.2 示例界面和语法：逐步法（BSTEP）

将一个样本分为具有症状“局部缺血”的患者（组类别“局部缺血”）和一个对照组（组类别“对照”）。在两个组中，分别提取不同定比数据（主要是血压参数），相互比较两个组中的这些变量。由于不清楚两组的哪些血压参数可能有多大程度上的差异，因此用逐步法建立了一个模型，其作用是根据自变量尽量完美地预测归属于“局部缺血”组或者“对照”组的概率。根据模型中引入的每个（如二元尺度）因子的变量，所建立的模型可以推导出对胜率的估

计。这个胜率说明了与对照组相比（由于编码为 0，“对照”是参考类别），一个人有多大的概率可以被归类到患者组“局部缺血”（胜率， $\text{Exp}(\mathbf{B})$ ）。根据模型中引入的定距变量，可以推导出影响权重（ \mathbf{B} ），通过针对每个患者的一个回归方程，这个影响权重可以表达出该患者归属于某个组的概率。

备注：或许事先还要指出的是，所使用的模数据是接近实际的，也就是说不是最佳的，这是为了在进行解释时能够指出某些需要注意的“陷阱”。

界面选择（示例）

在 SPSS 程序主界面选择以下菜单项：分析 → 回归 → 二元 Logistic。

把变量“组”拖入窗口“因变量”。将变量 ALTER、SYS_VT、SYS_VN、DIA_VT 和 DIA_VN 拖入窗口“协变量”。在“方法”一项下设置“向后：逻辑回归”逐步法。



子窗口“分类”：这个子窗口与本次分析无关。这个分析模型不含有分类变量。单击“继续”按钮。

子窗口“选项”：选定选项“分类图”、“Hosmer-Lemeshow 拟合优度统计量”和“ $\text{Exp}(\mathbf{B})$ 的置信区间”。在“显示”一项下选择，应在“每一步”模式下输出统计量和图形。在“逐步法概率”一项下可以给定一些标准，根据这些标准将自变量纳入方程（ $p=0.05$ ）或者从方程中删除（预设置： $p=0.10$ ）。在“分类阈值”一项下，（标准化地）给定 0.5 作为将病例分类的分割点值。在“迭代次数”一项下，用预设置的数值 20，则这个模型（只能）最多只能迭代 20 次。选中“将常数引入模型”选项，从而确定模型应包含一个截距（ b_0 ）。单击“继续”按钮。

子窗口“保存”：在这个例子中，不保存残差、高杠杆值或者预测值。单击“继续”按钮。

单击“确定”按钮开始计算。

语句：

```
LOGISTIC REGRESSION
  VARIABLES=gruppe
  /METHOD=BSTEP (LR) alter sys_vt sys_vn dia_vt dia_vn
  /PRINT=GOODFIT CI (95)
```

```
/CRITERIA PIN (.05) POUT (.10) ITERATE (20) CUT (.5)
/CLASSPLOT .
```

备注：LOGISTIC REGRESSION 命令调用逻辑回归方法。在“VARIABLES=”一项下，给出分别具有“局部缺血”（编码为 1）和“对照”（编码为 0，参考类别）水平的二元尺度因变量（在这里：组）。因变量可以是定量的或者二元的（理想情况下较短的）字符串变量。在/METHOD 一项下，在模型的效果之前确定选择的方法（进入、向前、向后）和一个统计量（条件、似然比、Wald）。在“选择方法”一项下，可以在下列组合中做出选择：“进入”、“向前条件”、“向前 LR”、“向前 Wald”、“向后条件”、“向后 LR”和“向后 Wald”（详情见下文）。在采用逐步法时，可以根据（条件）得分统计量（COND）、似然比统计量（LR）或者 Wald 统计量（WALD）进行剔除。通常（条件）得分统计量用于进入，Hosmer & Lemeshow（2000）、Menard（2001）和 SPSS（例如 V12，2003）的技术文档明确建议使用似然比统计量；当回归系数较大时，Wald 统计量会导致 II 类错误（尽管出现了现象，但是一次检验没有达到显著性）。因此，为了设定模型，可以成组或者单个地进入或者剔除带有针对分析方法和统计量的下列选项中的某一变量：ENTER “进入”（预设置）、FSTEP（COND）“向前条件”、FSTEP（LQ）“向前 LR”、FSTEP（WALD）“向前 Wald”、BSTEP（COND）“向后条件”、BSTEP（LQ）“向后 LR”和 BSTEP（WALD）“向后 Wald”。

ENTER（进入）：在一个步骤中进入所有变量（预设置为在步骤 1 后，ENTER 立即停止）。只有当预测变量的数量和判别效率是已知的或者至少是固定设置时，例如，在计算有序逻辑回归时，才能使用强制进入法。如果判别潜力不明，或者要测定一个只含有很少变量的有效预测模型时，则应使用逐步法（例如，通过 FSTEP 或者 BSTEP）。

FSTEP（向前步进）：根据 FSTEP，依次检验变量或交互效应是否可以进入模型。得分统计量显著性水平最低（小于 PIN）的变量进入模型。然后检验模型中现有的所有变量是否可以剔除。删除在所对应统计量（可选择似然比、Wald、条件）中显著性值最高的变量（超过 POUT）。重新检验剩下的模型是否可以剔除。只要不再有变量满足剔除标准，就检验协变量是否可以进入模型。不断重复这个过程，直到所有变量满足进入或者剔除标准为止。

BSTEP（向后步进）：第一步根据 BSTEP 一次性将所有变量或者交互效应进入模型，然后依次检验是否可以剔除。逐步剔除或者进入基本与 FSTEP 的流程一致，不断重复这个过程，直到所有变量符合进入或者剔除标准为止。对于 FSTEP 和 BSTEP，可以选择检验统计量，然后在括号中给出（COND：条件、WALD：Wald、LR：似然比）。

根据这个方法给定效应，在这种情况下就是定距变量，在本个案中即“alter”、“sys_vt”等。至少给定一个预测变量，用 a*b 格式定义交互作用。如果将分类（离散分级）因子（例如，变量 DIAGNOSE1 和 DIAGNOSE2）设定为协变量，则必须将其额外地录入单独的 CONTRAST 语句，例如：

```
/METHOD= BSTEP (LR) alter sys_vt sys_vn dia_vt dia_vn
               diagnoses1 diagnoses2
/CONTRAST (diagnoses1) = indicator
/CONTRAST (diagnoses2) = indicator
```

可以同时给定多个/METHOD 命令行，例如，用于有序逻辑回归。这里所介绍的示例借助

于选择方法“向后步进”和 LR 统计量，调查了给定的五个定距变量中的哪一个可能对因变量 GRUPPE（组）施加影响。

在 PRINT 一项下确定输出结果。利用 CI（95，预设置）定义胜率（Exp(B)）的置信区间为 95%；在 CI 一项下，可以给定 1 到 99 的整数。如果测定的置信区间（如 95%）包含了数值 1，则可以得出结论：模型中的自变量有 95% 的概率对因变量各个水平之间的比值（以及胜率）不产生影响。例如，GOODFIT 调用 Hosmer-Lemeshow 拟合优度统计量（GOODFIT）。向每个可用的输出结果询问关键词 ALL。SPSS 标准化地输出内容丰富的表，尤其是对于多项模型的步进分析，这些表可以涵盖较广的范围。通过 SUMMARY 可以输出一个所有等级的综合表，作为步进输出的代替。

在 CRITERIA 一项下，可以将迭代计算回归模型的参数传递给 SPSS。需要给定的变量主要取决于在 METHOD= 一项下指定了哪种方法。只要达到目标标准（例如，ITERATE、BCON 或 LCON），迭代就结束。利用 PIN 和 POUT 确定，根据哪些参数决定变量进入模型还是从模型中剔除。通过 PIN（0.05），设定了决定一个变量是否可以进入模型的数值。如果一个变量的得分统计量的概率小于进入值，则这个变量进入模型。给定的进入值（PIN）越大，变量就越容易进入模型。SPSS 中预设置的 0.05 被认为是相对保守的，为了使潜在的有关预测变量进入，最多可以接受 0.20。利用 POUT（预设置 0.10）指定一个数值，根据这个数值衡量是否从模型中剔除一个变量。如果概率超过剔除值，则根据有条件的 LR 或者 Wald 统计量将这个变量剔除。给定的剔除值（POUT）越大，变量就越容易留在模型中。进入值必须小于剔除值。ITERATE（20）设定一个为正整数的迭代最大次数，例如，在这里是 20。通过迭代的方法，估计最大似然系数。如果达到迭代的最大次数，则在达到收敛之前停止迭代。

CUT（0.5）确定了针对所估计概率的分割点值，根据这个概率，将所测定变量的预测值分配给某个组。CUT 主要影响预测组、分类表和分类图，其标准值为 0.5。

在每一个步骤中，CLASSPLOT 命令都以直方图的形式给定了二元因变量观察（actual）值和预测（predicted）值的分类。

根据 SPSS 语法，在上述的语句示例中还可以设置更多的输出结果。为了计算逻辑回归，可以根据自己的要求提供很多选项（例如，CASEWISE、MISSING 或 SAVE），进一步区分示例中所介绍的 LOGISTIC REGRESSION 程序。详细情况可以参见 SPSS 语法文档和统计学专业文献。

3.2.3 输出结果和解释

逻辑回归

在这个标题后面是所调用的二元逻辑回归的输出结果。

个案处理总结

未加权个案 ^a		N	百分比
选择的个案	纳入分析的个案	160	100.0
	缺失个案	0	0.0
	总数	160	100.0
未选择的个案		0	0.0
总数		160	100.0

a. 如果进行了加权，则可以在分类表中获取个案总数。

针对每次分析，表“个案处理总结”都给出了个案总数（例如， $N=160$ ）、纳入分析的个案数量（例如， $N=160$ ）和缺失个案的数量（例如， $N=0$ ）。

因变量编码

原始值	内部值
对照	0
局部缺血	1

表“因变量编码”输出了因变量的内部编码。如果没有给出这些数量，则无法明确地解释（尤其是对于分类预测变量）所输出的参数（回归系数，胜率）。在 LOGISTIC REGRESSION 中，所测定的统计量始终是针对事件“局部缺血”（1）相比于参考类别（“对照”）。在本例中，系数为正表示目标事件“局部缺血”的发生概率较大。如果结果是针对作为目标事件的“对照”，则应将因变量重新编码，从而将“对照”编码为 1（语句中的替代选择，但在较旧版本中没有）。尤其是对于字符串变量，内部编码与数据集的编码有所不同。如果有分类变量，则也输出参数编码。各个预测变量应相互不相关。为此，从 SPSS 中可以调用一个相应的相关图。

针对每个迭代步骤，都显示进入的和删除的变量。在迭代记录中，迭代步骤的表现始终从步骤 0 开始。但是，所显示的内容始终取决于所选择的方法。回忆一下，本例使用了“向后步进”法和 LR 统计量，这表示，初始组块是基于没有自变量的一个模型。利用表“分类表”、“方程中的变量”和“不在方程中的变量”展现了初始组块。

初始组块

分类表^{a,b}

观察			预测		
			组		正确百分比
			对照	局部缺血	
步骤 0	组	对照	0	62	0.0
		局部缺血	0	98	100.0
总百分比					61.3

a. 常数已输入模型。

b. 分割点值为 0.500。

“分类表”原则上是因变量观察值和预测值的一个交叉列表，从而可以对模型的性能进行

初步估计。错误的预测值越少，模型拟合优度越高。如果观察值超过分界值（预设置为 0.5），则因变量的预测值视为 1，否则就视为 0。从左上到右下的对角线上的数值表示正确的预测（例如，0 和 98），在对角线旁边大于 0 的数值表示不正确的预测（例如，62）。如果预测 100% 正确，则对角线旁边的单元格只含有 0。在没有考虑定量预测变量的情况下，作为示例的模型做出的预测大约 61% 是正确的。

方程中的变量

	回归系数 B	标准误差	Wald	Df	Sig.	Exp(B)
步骤 0 常数	0.458	0.162	7.960	1	0.005	1.581

表“方程中的变量”和“不在方程中的变量”反映了在步骤 0 时模型的参数。在表“方程中的变量”中，唯一的参数是常数。对于预测值，主要显示得分（“数值”）和显著性（“Sig”）。

不在方程中的变量

		数值	df	显著性
步骤 0	变量 年龄	3.423	1	0.064
	sys_vt	3.325	1	0.068
	sys_vn	0.146	1	0.702
	dia_vt	0.505	1	0.477
	dia_vn	0.430	1	0.512
	总统计量	15.878	5	0.007

表“不在方程中的变量”中显示了所选择的（但是还没有进入，因为是步骤“0”）变量和相应的得分统计量。只要有一个数值的显著性（“Sig”）低于 0.05，则这个数值就视为显著。根据所选择的方法不同，对显著性应予以不同的解释：在采用向后法时，显著性值大（如本例中的年龄）的变量首先被模型保留；在采用向前法时，显著性值小（如本例中的年龄）的变量首先进入模型。在采用其他任何一种方法时，都是显著性值最低的变量被添加到模型中。在每个步骤中都测定“总统计量”，并检验零假设是否成立，即不在回归方程中的每个变量都具有系数 0。在这种情况下，“总统计量”具有显著性。不在回归方程中的任何一个变量都具有不等于 0 的系数。

组块 1：方法 = 向后法（似然比）

这个标题给定了由用户编排的变量块的编号（“1”）、所选择的方法（“向后步进”）以及所选择的统计量（“似然比”）。在这些组块中具体包括哪些变量，将在后文中予以阐述。

模型系数综合检验

	卡方	df	显著性
步骤 1 步骤	16.819	5	0.005
组块	16.819	5	0.005
模型	16.819	5	0.005

续表

	卡方	df	显著性
步骤 2 ^a 步骤	-0.213	1	0.645
组块	16.607	4	0.002
模型	16.607	4	0.002
步骤 3 ^a 步骤	-0.179	1	0.672
组块	16.427	3	0.001
模型	16.427	3	0.001

a. 卡方的负值表明，先前级别的卡方已经降低。

从表“模型系数综合检验”中可以查阅相应步骤中模型性能的量度。“卡方”给出了 2*对数似然值相较于前一个步骤、组块或者模型的变化。p 值（“Sig”）的解释取决于所选择的方法。在使用“向后步进”法（详情见下文）时，先后剔除变量。只要显著性的变化足够大，例如，超过 0.10，就一直剔除变量。在本例中，在步骤 3 时的显著性与步骤 2 时的相比，已经只有 0.027 的变化幅度了（ $0.027=0.672-0.645$ ）。因此不再执行第 4 个步骤，因为显著性变化的幅度已经低于 0.10 了。

模型总结

步骤	2*对数似然值	Cox & Snell R ²	Nagelkerke R ²
1	196.818 ^a	0.100	0.135
2	197.031 ^a	0.099	0.134
3	197.210 ^a	0.098	0.132

a. 估计过程在第 4 次迭代时结束，因为参数估计值的变化幅度小于 0.001。

从表“模型总结”中可以查阅在每个步骤时测定的 2*对数似然值和拟 R²统计量（Cox & Snell R², Nagelkerke R²）。R²统计量近似地测量由模型解释的、因变量中的变异分量。R²统计量越大（只有在 Nagelkerke R²时最大值=1.0），解释的方差分量越大。在本例中，R²统计量很小。模型只能解释很小的方差分量，根据 Nagelkerke 的著作大约只有 13%。

Hosmer-Lemeshow 检验

步骤	卡方	df	显著性
1	4.905	8	0.768
2	4.529	8	0.807
3	7.520	8	0.482

针对每个迭代步骤，都可以从表“Hosmer-Lemeshow 检验”中得到模型拟合优度的一个量度，同时检验零假设是否成立，即模型已经恰当地拟合了数据。只要显著性小于 0.05，那么模型就没有适当地拟合。如果有逐个案的数据，也就是如果有很多预测变量或者预测变量是连续的，则 Hosmer-Lemeshow 拟合检验特别适合确定整个的模型拟合优度。Hosmer-Lemeshow 检验是一种经过改良的皮尔逊卡方检验，是基于分布在 10 个同样大小的组中的期望概率。期望频率的数量应符合卡方检验的标准，如果没有占满或者空白的单元格（见下文）过多，则检验的结果可能不可靠。

Hosmer-Lemeshow 检验列联表

		组=对照		组=局部缺血		总数
		观察	期望	观察	期望	
步骤 1	1	9	10.569	7	5.431	16
	2	8	8.497	7	6.503	15
	3	10	8.490	7	8.510	17
	4	6	7.184	10	8.816	16
	5	9	6.871	8	10.129	17
	6	4	5.681	12	10.319	16
	7	7	5.237	10	11.763	17
	8	4	4.277	13	12.723	17
	9	4	3.433	12	12.567	16
	10	1	1.762	12	11.238	13
步骤 2	1	9	10.672	7	5.328	16
	2	11	9.421	6	7.579	17
	3	6	7.792	10	8.028	16
	4	8	7.417	9	9.583	17
	5	9	6.707	8	10.293	17
	6	5	5.580	11	10.420	16
	7	6	4.885	10	11.115	16
	8	4	4.414	13	12.586	17
	9	3	3.361	13	12.639	16
	10	1	1.572	11	10.428	12
步骤 3	1	8	11.200	9	5.800	17
	2	9	8.297	6	6.703	15
	3	12	8.478	5	8.522	17
	4	6	7.184	10	8.816	16
	5	6	6.765	11	10.235	17
	6	6	5.417	10	10.583	16
	7	7	5.224	10	11.776	17
	8	4	4.412	13	12.583	17
	9	3	3.597	14	13.403	17
	10	1	1.426	10	9.574	11

在迭代步骤时，可以从“Hosmer-Lemeshow 检验列联表”中查阅用于计算 Hosmer-Lemeshow 检验的基本数据。Hosmer-Lemeshow 检验是基于分布在 10 个同样大小的组中的期望概率。针对二元因变量的每个水平，这个表格给出了观察到的和模型预计的个案数量。在步骤 3，在“对照”水平中很低的期望值进入子组 8、9 和 10，两个很低的观察频率进入子组 8 和 9。

分类表^a

观察		预测		
		组		正确百分比
		对照	局部缺血	
步骤 1 组	对照	21	41	33.9
	局部缺血	16	82	83.7
	总百分比			64.4
步骤 2 组	对照	23	39	37.1
	局部缺血	16	82	83.7
	总百分比			65.6
步骤 3 组	对照	23	39	37.1
	局部缺血	15	83	84.7
	总百分比			66.3

a 分割点值为 0.500。

“分类表”是因变量观察值和预测值的一个交叉列表。与步骤 0（见上文，正确预测占 61.3%）相反，到步骤 3 时，由于引入了各种预测变量，预测正确率达到了约 66%。通过添加定量预测变量，模型性能改善了大约 5%。引人注目的是，“局部缺血”组（大约 84%）可以比“对照”组（大约 37%）有更高精确度的分类。总而言之，这个模型性能不是很好。

方程中的变量

		回归系数 B	标准误差	Wald	df	Sig.	Exp(B)	EXP(B)的 95%置信区间	
								下限值	上限值
步骤 1 ^a	alter	-0.035	0.014	6.293	1	0.012	0.966	0.939	0.992
	sys_vt	0.215	0.065	10.946	1	0.001	1.239	1.091	1.408
	sys_vn	-0.029	0.045	0.395	1	0.530	0.972	0.889	1.062
	dia_vt	-0.213	0.095	5.079	1	0.024	0.808	0.671	0.973
	dia_vn	0.028	0.061	0.213	1	0.645	1.028	0.913	1.158
	Konstante	2.103	1.014	4.305	1	0.038	8.192		
步骤 2 ^a	alter	-0.035	0.014	6.218	1	0.013	0.966	0.940	0.993
	sys_vt	0.207	0.063	10.928	1	0.001	1.231	1.088	1.392
	sys_vn	-0.013	0.031	0.181	1	0.671	0.987	0.930	1.048
	dia_vt	-0.200	0.089	4.987	1	0.026	0.819	0.687	0.976
	Konstante	2.126	1.014	4.396	1	0.036	8.379		
步骤 3 ^a	alter	-0.035	0.014	6.439	1	0.011	0.965	0.939	0.992
	sys_vt	0.204	0.062	10.838	1	0.001	1.227	1.086	1.385
	dia_vt	-0.202	0.089	5.158	1	0.023	0.817	0.686	0.973
	Konstante	2.077	1.004	4.280	1	0.039	7.980		

a. 在步骤 1 输入的变量：alter、sys_vt、sys_vn、dia_vt 和 dia_vn。

表“方程中的变量”反映了在每个步骤中模型的变量和参数。在步骤 0 时，这些数据在表“不在方程中的变量”中，本书不阐述逐步输出的表“不在方程中的变量”和“删除项时的建模”。表“不在方程中的变量”的步骤 0 已经给出了得分统计量（“数值”），但在表“方程中的变量”中给出的却是 Wald 统计量。在模型中只含有定比变量。对其参数的解释与分类变量的解释只有细微区别（参见 Menard, 2001）。为了可以可靠地解释这个表格，至少因变量的编码必须是已知的（对此参见表“分类变量的编码”）。在本例中，测定的统计量始终是指“局部缺血”组相比于参考类别“对照”。

分类变量的输出结果是根据模型因变量具有两个还是超过两个变量类别来区分的。对于具有超过两个变量类别的变量，首先输出变量的 Wald 统计量（也就是说，除了参考类别的回归系数之外，是否所有的回归系数都等于 0），然后输出代表各个独立类别的系数的 Wald 统计量。对于只有两个变量类别的变量，输出参考类别的 Wald 统计量。在表的括号中，给出了各自的类别或类别编码。在有数据缺失的情况下，给定的变量类别不总是与编码一致，因此建议仔细检验类别等级的完整性。最后一个类别每次都是冗余的，这是根据其他分组得出的。具体而言，这就表示无须明确的编码，只利用常数，也就是不用回归系数就可以计算最后一个类别的概率。

针对预测变量给出了大量参数。“B”是估计的非标准化回归系数，对此显示 B 的标准误差。B 除以其标准误差的平方就得出了 Wald 统计量。如果 Wald 统计量（“Wald”）小于 0.05，则相应的参数统计上是显著的，即该参数区别于 0，相应的变量对于模型是有用的。

对于回归系数 B，正负号和数值都很重要。正数的回归系数表示，随着预测值的升高，各自参考类别的替代类别的概率也在升高；而负数的回归系数（例如，“SYS_VT”）则表示，随着预测值的升高，参考类别的概率随之升高。重要的是，SPSS 输出非标准化系数，因此对这种非标准化系数的估计无论如何是不可靠的。在这里， $\text{Exp}(B)$ 可以起到指引方向的作用： $\text{Exp}(B)$ 的数值与 1 的偏差越大，相对来看回归系数也就越大（附录中指出了与回归系数解释相关的一些特殊性）。

“ $\text{Exp}(B)$ ”是胜率（不能与相对风险的量度混淆，例如，参见 Schendera 著作第 12 章关于表格分析的部分，2004），说明了当预测变量升高一个单位时，所预测的胜率变化幅度。 $\text{Exp}(B) > 1$ 表明因变量的比值上升； $\text{Exp}(B) < 1$ 则表明因变量的比值下降。对于 $\text{Exp}(B)$ 而言，重要的是所调用的置信区间是否包含 1。置信区间的数值超过 1 越多，相应变量的影响也就越大。可以十分简便地通过 $\text{Exp}(B)$ 看出非标准化定量预测变量的相对意义（B 是非标准化的，可能有很大的误导作用）。通常，只解释显著性预测变量的 $\text{Exp}(B)$ 。概率可以解释为根据其他预测变量做了调整。在表格下面，给出了初始模型（步骤 0）的变量选择。

在第 3 步，例如，胜率 $\text{Exp}(B) = 1.227$ 表示，当预测变量“SYS_VT”升高一个单位时，因变量的比值随之增大。因此，如果因变量中的比值（编码为 0 和 1）先前是 1:1，则预测变量“SYS_VT”增大一个单位会导致比值变成 1:1.227。在预测变量升高一个单位时，编码为“1”的因变量类别（“局部缺血”）与“对照”组相比，概率升高了 22.7%，或者说达到原先的 1.227 倍。对于预测变量“年龄”， $\text{Exp}(B) = 0.965$ 表示，随着预测变量“年龄”升高一个单位，因变量类别“1”的发生概率随之减小。因此，如果因变量中的比值先前是 1:1，则预测变量“年龄”升高一个单位会导致比值变成 1:0.965。在预测变量升高一个单位时，编码为

同样，预测的概率大部分是不精确的。在大约 0.3 到 0.5 的值域（“对照”， <0.5 ）分布了预测不正确的“局部缺血”病例，“对照”病例不正确地分布在 0.5 到 0.85 的“局部缺血”值域内。分类图明确地展示出，模型不能正确地预测所观察的病例，因此不能总是正确地将“局部缺血”病例分类为“局部缺血”。换言之，很多“局部缺血”病例被预测为属于“对照”组，也就是预测为假阳性，这除了表明预测效率不足之外，还可以理解为存在离群值和模型拟合优度不佳。

总结

一个样本分为具有“局部缺血”病症的患者组和对照组。从两个组中，提取了不同定比数据（主要是血压参数）。由于不清楚两组的哪些血压参数可能有多大程度的差异，这里利用逐步法建立了一个模型，其作用是根据自变量尽量完美地预测某个病例归属于“局部缺血”组或者“对照组”的概率。所建立的模型在第 3 个，也就是最后一个步骤是符合要求的（ $\text{Chi}^2=16.427$ ， $p=0.0001$ ，模型系数综合检验），但是只能解释很小的方差分量（Nagelkerke: 大约 13%）。根据 Hosmer-Lemeshow 检验，模型没有相应地拟合数据（ $p=0.031$ ），由于单元格没有占满，因此这个检验的可靠性是值得质疑的。在步骤 3，这个模型在引入变量 ALTER、SY_VT 和 DIA_VT 后，得出约 66% 正确分类的病例。但是，“局部缺血”组的分类（大约 84%）明显比“对照”组的分类（大约 37%）好很多。测定的胜率（“Exp(B)”）表明，胜率在总体上只发生了很小的变动。从分类图可以看出，主要是“局部缺血”个案被预测为可能属于“对照”组，也就是预测为错误的正数，这可以理解为预测效率不佳。

3.2.4 示例和语法：直接法 ENTER

将一个样本（因变量“组”）分为病例（组类别“病例”）和对照（组类别“对照”）。从这两个组中，分别提取不同定比型实验室参数。根据临床经验，用这些实验室参数组成一个对变量（LABOR1、LABOR2 和 LABOR3）的验证数据集。应检验变量的组成，即这些变量作为对照，是否可以和猜测的一样明确地预测出一个病例归属于“病例”组还是“对照”组。在这种情况下，利用直接法检验模型。

语句：

```
LOGISTIC REGRESSION
  VARIABLES=gruppe
  /METHOD=ENTER labor1 labor2 labor3
  /CLASSPLOT
  /PRINT=GOODFIT
  /CRITERIA PIN(.05) POUT(.10) ITERATE(20) CUT(.5).
```

备注：本语句在很大程度上与前面的示例一样。因此，在这里只指出与分析相关的特殊性，即在 /METHOD 一项下，利用 ENTER 命令调用“进入”法。在使用“进入”法时，所有变量只通过一个步骤就一次性进入。因此在步骤 1 后，进入法立即停止运行。只有当预测变量的数量和判别效率是已知的或者至少是固定设置的时，才能使用进入法。如果判别潜力不明或者要测定一个所含变量尽量少的有效预测模型，则应使用逐步法（通过 FSTEP 或者 BSTEP）。根据这个方法给定效应，在这种情况下就是定距变量（在这里是 LABOR1 等）。这里介绍的示例通过“进入”法调查了三个变量 LABOR1、LABOR2 和 LABOR3 是否可以明确地归类到病

例个案和对照个案。

3.2.5 输出结果和解释

逻辑回归

在这个标题后面，是一个二元逻辑回归的输出结果。

个案处理总结

未加权个案 ^a		N	百分比
选择的个案	纳入分析的个案	157	98.1
	缺失个案	3	1.9
	总数	160	100.0
未选择的个案		0	0.0
总数		160	100.0

a. 如果进行了加权，则可以在分类表中获取个案总数。

针对分析，表“个案处理总结”给出了个案的总数（ $N=160$ ）、有效个案的数量（ $N=160$ ）和缺失个案的数量（ $N=0$ ）。

因变量编码

原始值	内部值
对照	0
病例	1

表“因变量编码”输出了因变量的内部 0/1 编码。“对照”组是参考类别，所测定的统计量始终是针对“病例”组。如果结果是针对作为目标事件的“对照”组，则应将因变量重新编码，或者通过语句切换参考类别。

针对每个迭代步骤，都显示进入的和删除的变量。在迭代记录中，当采用进入法时，迭代步骤的表现始终从步骤 0 开始。

初始组块

分类表^{a,b}

观察			预测		
			组		正确百分比
			对照	病例	
步骤 0	组	对照	0	62	0.0
		病例	0	95	100.0
总百分比					60.5

a. 常数已输入模型。

b. 分割点值为 0.500。

“分类表”可以对没有预测变量的模型的性能进行初步估计。示例中的模型在步骤 0 时做出大约 60% 的正确预测。需要注意的是，没有同样正确地对病例个案和对照个案做出预测

(0 vs 100%)。

方程中的变量

	回归系数 B	标准误差	Wald	df	Sig.	Exp(B)
步骤 0 常数	0.427	0.163	6.832	1	0.009	1.532

表“方程中的变量”和“不在方程中的变量”反映了在步骤 0 时模型的参数。在表“方程中的变量”中，唯一的参数是常数。对于预测值，主要显示得分（“数值”）和显著性（“Sig”）。

不在方程中的变量

	数值	df	显著性
步骤 0 变量 labor1	4.577	1	0.032
labor2	10.530	1	0.001
labor3	3.399	1	0.065
总统计量	25.875	3	0.000

在表“不在方程中的变量”中，显示了所选择的（但是还没有进入的）变量及其得分统计量（“数值”）。“总统计量”具有显著性。任何一个（还）没有进入模型的变量都具有不等于 0 的系数。

组块 1：方法 = 进入

这个标题给出了所组成的变量块的编号（“1”）以及所选择的方法（“进入”）。

模型系数综合检验

	卡方	df	显著性
步骤 1 步骤	28.494	3	0.000
组块	28.494	3	0.000
模型	28.494	3	0.000

从表“模型系数综合检验”中，可以查阅步骤 1 时（也就是所选择的预测变量完全进入模型后）模型性能的量度。“卡方”给出了 2*对数似然值相较于前一个步骤、组块或者模型的变化。p 值（“Sig”）的解释取决于所选择的方法。例如，如果要采用进入法使变量进入，则只有当 p 值的变化的显著性足够小，例如，低于 0.05 时，才能使这些变量进入模型。在本例中，使变量集进入模型是有意义的，因为相较于前一个步骤、组块或者模型，2*对数似然值变化的显著性始终低于 0.05，或者不超过 0.05。

模型总结

步骤	2*对数似然值	Cox & Snell R ²	Nagelkerke R ²
1	182.166 ^a	0.166	0.225

a. 估计过程在第 5 次迭代时结束，因为参数估计值的变化幅度小于 0.001。

从表“模型总结”中可以查阅在每个步骤时测定的 2*对数似然值和拟 R²统计量（Cox & Snell R², Nagelkerke R²）。模型只能解释很小的方差分量（根据 Nagelkerke 的著作，大约只

有 22.5%)。

Hosmer-Lemeshow 检验

步骤	卡方	df	显著性
1	17.128	8	0.029

根据表“Hosmer-Lemeshow 检验”可知，模型没有恰当地拟合数据 ($p=0.029$)。但是检验的可靠性是值得质疑的（见下文）。

Hosmer-Lemeshow 检验列联表

		组=对照		组=病例		总数
		观察	期望	观察	期望	
步骤	1	15	12.590	2	4.410	17
1	2	11	9.990	5	6.010	16
	3	7	8.443	9	7.557	16
	4	3	7.983	14	9.017	17
	5	8	6.950	9	10.050	17
	6	3	5.920	14	11.080	17
	7	8	4.113	8	11.887	16
	8	4	3.184	12	12.816	16
	9	3	2.203	13	13.797	16
	10	0	0.625	9	8.375	9

针对二元因变量的每个变量类别，表“Hosmer-Lemeshow 检验列联表”给出了观察到的和模型预计的个案数量。在“对照”组中，很低的期望值进入子组 8、9 和 10。在“病例”组中，很低的期望值进入子组 1。

分类表^a

观察			预测		
			组		正确百分比
			对照	病例	
步骤 1	组	对照	33	29	53.2
		病例	16	79	83.2
总百分比					71.3

a. 分割点值为 0.500。

“分类表”表明，包括三个参数 LABOR1、LABOR2 和 LABOR3 在内的模型以大约 70% 的正确率预测了归属于哪个组。相比于没有预测变量的模型，这个模型的性能提高了 10%。此外，缓解了对病例个案（100%~83%）和对照个案（0%~53%）的预测大幅度不一致的现象。

成的变量集 LABOR1、LABOR2 和 LABOR3 预测效率不佳。

此外，SPSS 还可以显示出模型的卡方、改进的卡方、变量之间相关性的图形、观察组和预测概率、卡方残差和对数似然值（如果从模型中已经删除了项目）。

3.2.6 补充说明逻辑回归的理论检验 vs 诊断：模型拟合优度 vs 预测效率

二元逻辑回归和多项逻辑回归原则上遵循两个原则，首先是检验一个理论，然后是检验个案正确分类的准确性。经常是（在理想情况下通过理论推导测定并经过检验的）模型拟合优度和与其同时产生的预测精度（预测效率）相互一致就够了。例如，可以通过一个交叉列表确保实现的预测精度。

但是，模型拟合优度和预测精度不总是相互一致的。具有良好参数（例如，卡方、McFadden R^2 ）的模型很可能有十分糟糕的预测精度（预测效率），相反，拟合优度量度很差的模型可能达到相当好的预测效率。这是为什么呢？

其原因是，逻辑回归模型的拟合优度的量度是基于对数似然值（-2LL），但是没有考虑在本模型中正确和不正确分类的个案之间的比值。因此，模型拟合优度的量度不能说明模型的预测精度。这个预测效率（也就是正确或不正确分类的个案的比值）不能用通常的表格相关性量度测定（例如， ϕ 系数，Goodman & Kruskal 等人著作，Menard, 2001）。如今，在 SPSS 中没有实现由 Menard（2001，参见其公式）建议使用的逻辑回归分类模型的量度。

3.2.7 二元逻辑回归的前提条件

1. 逻辑回归假设了（至少）一个自变量（ X ）和因变量（ Y ）之间的因果模型。根据逻辑，从一开始就排除了伪回归，例如，性别对收入的预测。在模型中只给出了重要的变量，不重要的变量也应删除。
2. 成对的测量值 x_i 和 y_i 必须属于同一个对象。换言之：所调查的特征是从一个样本的同一个元素中提取的。
3. 理想情况下，自变量和因变量相关紧密。
4. 因变量。在二元逻辑回归中，因变量是二元的；在多项逻辑回归中，因变量超过两个变量类别。在二元逻辑回归时，重要的是检验令人感兴趣的目标事件是否（频率足够地）发生，以及目标事件的频率究竟是符合总体，还是呈现出不均衡。如果目标事件很少，则通常假阴性的成本要超过假阳性的成本，并且分界值远远低于 0.5。
5. 缺失数据（缺失值）。尤其是对于预测模型，缺失数据可能导致产生问题。预测模型的理想条件是不缺失任何数据。如果数据是完全随机缺失的，则具体的缺失程度决定了分析时还留有多少百分比的数据，这可能会导致出现问题。如果通过合理的思考，发现缺失值以某种方式与目标变量相关，那么从模型中剔除了这些缺失值，模型的解释和建模就会产生问题。例如，（a）从建模角度通过一个指示缺失值的指标和（b）从重建角度分析缺失数值（Missing Value Analysis），可以将缺失数据重新引入模型，但是只能在这个前提条件下：这些缺失数据的编码、重建和模型集成是合理和可追溯的。如果缺失值集中在一个变量上，则或许也可以从分析中剔除这些缺失值。

6. 逻辑回归假设了因变量和自变量之间的非线性函数，以及胜率对数的线性，即连续预测变量和因变量胜率对数之间的线性关联，这些都可以想象成一个线性的散点图。但是，目前在 SPSS 中还不能方便地输出 Logit 图。一种检验胜率对数线性假设的简便方法是，用每个连续预测变量及其固有算法之间的交互作用项补充初始的回归模型（称之为 Box-Tidwell 转换）。如果这些交互作用项其中的一个呈现出显著性（相对于其他交互作用，参见 Jaccard, 2000），则这个模型就违反了胜率对数的线性假设，以及单调关联的假设（关于忽视非单调关联的风险请参见 Böhning 著作，1998，第 6 章）。这个方法的缺点是，无法识别轻微偏离线性的现象，并且在达到显著性时无法反映出非线性的形状。Menard (2001) 介绍了其他检验方法。非线性不能与非叠加性相混淆。
7. 模型的叠加。如果因变量发生的、相当于自变量、一个单位的变化取决于自变量的数值，则出现非线性。与此相比，如果因变量相对于其中一个其他自变量的数值发生了达到自变量一个单位的变化，则呈现非叠加性。例如，可以通过检验是否存在可信的或者理论上可能有的所有交互作用，就可以检验模型的叠加性。后一种方法只适用于相对简单的模型。
8. 散布问题（超散布性或超聚集性，又称 Overdispersion 或 Underdispersion）。对于正确设定的模型，模型拟合优度（皮尔逊、离差）量度除以自由度的数量应得出一个在 1 左右的数值。若这个数值远远超过 1，则表明存在超散布性；低于 1 则表明是很少出现的超聚集性。散布问题表明误差也许不是呈现二元分布，并且散布问题是导致产生错误的主要标准误差。在分析实践中经常碰到的超散布性，是由于模型缺失重要的预测变量、或必须转换这些预测变量、或存在离群值而造成的。通过将协方差矩阵改变尺度可以对离散进行校正，但是只有在检验和排除了其他错误源之后才能实施。
9. 参考类别。参考类别对所测定结果的尺度和方向具有决定性预测。例如，胜率编码为 1 时可能数值达到 3，但是编码为 0 时可能达到 0.33。例如，对于二元因变量（B）的系数，正负号发生改变。SPSS 过程命令 LOGISTIC REGRESSION 始终选择因变量的第一个或者最低的变量类别作为参考类别（在发生事件时编码为 1）。SPSS 过程命令 NOMREG 标准化地选择因变量的最后一个或者最高的变量类别作为参考类别（但是从 SPSS 12 版本开始，可以单独给定参考类别）。其他作者、分析师或者软件在有些情况下使用了其他的参考类别。请检查（自动）选择的参考类别是否符合评估目的，否则就要将因变量重新编码。在临床或者流行病学研究中适用的规则是，始终将病例个案（暴露，事件）编码为“1”，始终将对照个案（不暴露，事件不发生）编码为“0”。更多信息，请参见多项逻辑回归的前提条件。
10. 自变量。预测变量应相互不相关（消除多重共线性）。预测变量之间的任何相关（例如 >0.80 ）都是多重共线性的迹象。通过允差检验，或者参数估计值的明显很高的标准误差（非标准化： >2 ，标准化： >1 ）显示出存在多重共线性。通过对同一个模型计算线性回归，可以测定允差量度（因为只测定预测变量之间关联的允差量度，因此允许使用这种处理方法，因变量在此无关紧要）。是否可以消除多重共线性和在多大程度上可以消除多重共线性，除了相关预测变量的数量和关联性之外，主要取决于错误出现在研究过程的哪个地方：理论构建、具体实施或者数据搜集。“如果发现了多

重共线性，具体怎么处理更像是一门艺术，而不是科学”（Menard, 2001, 80、也可参见 Pedhazur, 1982², 247）。

11. 个案。对于因变量的每个变量类别，应至少是 $N=25$ 。进入模型的预测变量越多，或者模型幂次越大，则需要的个案越多。Hosmer & Lemeshow（2000, 339-346）提出了一个公式，除了样本量之外还能给定模型的幂次和检验方向。如果多个预测变量等级的组合可能导致产生很多空白单元格，则既可以从模型中剔除不重要的预测变量，又可以将预测变量等级合并起来。明显很高的参数预测变量或者标准误差，以及分割理想的分类图（如完全分离或准完全分离），表明可能存在数据问题。为了保证数据完整性，应小心地将预测变量等级合并起来。
12. 每个参数的事件。仅仅有个案的数量是不够的，应保证目标事件（因变量）的两个变量类别中比较少见的一个，尤其是在有很多协变量时应以足够大的频率发生。作为经验法则，Hosmer & Lemeshow（2000, 346-347）建议针对有些对称的分布采用至少 $N=10 * n$ 个协变量（对于目标事件的不对称变量类别，情况更为复杂）。结果可能是只针对感兴趣的目标事件抽取个案、普遍地升高个案数量 N 以及（或者）减少协变量数量。
13. Hosmer-Lemeshow 检验（拟合优度检验，goodness-of-fit-Test）。Hosmer-Lemeshow 检验是一种经过改良的皮尔逊卡方检验，基于分布在 10 个同样大小的组中的期望概率（参见列联表，因此也可以称之为“deciles of risk”检验）。应通过卡方检验来检查期望频率是否恰当。期望频率的数量应至少 >2 ，只有对于 20% 的单元格才应 <5 ，并且在理想情况下不等于 0。如果单元格是没有占满的，并且用户不想接受较低的模型性能，则可以从模型中剔除不重要的预测变量，或者将预测变量等级合并起来。在解释拟合优度检验时，应注意检验方向和样本量：不显著的结果通常表示模型拟合优度良好。但是由于当样本很大时，即使最小的区别也会呈现出显著性，因此显著的结果并不一定表示模型拟合优度不佳。对于个案数据，Hosmer-Lemeshow 检验应始终优先于偏差或者皮尔逊这两个拟合优度量度。偏差或者皮尔逊不适合带有定比预测变量的模型。
14. 残差（误差）。LOGISTIC REGRESSION 中的残差是基于个案数据；对于成组的皮尔逊残差，可以换为采用 SPSS 过程命令 NOMREG。相互独立地对所调查的特征进行取样。因此，残差应是相互独立的。残差围绕着 0 随机散布，具有恒定的方差（同方差性），呈现二元分布（只有在样本很大时才呈现正态分布），既不相互相关，又不与预测变量相关。残差的分布取决于样本量。对于小的样本，违背这个假设被认为比大的样本要严重得多（前提条件是中央极限定理许可）。关于误差的独立性，可以通过消除自相关予以保证。对于可能通过“时间”因子而相互关联的变量（如在重复测量设计时），可以参见 Hosmer & Lemeshow（2000, 第 8.3 节）著作中阐述的逻辑方法的特定前提条件和用途。例如，SPSS 过程命令 GENLIN 可以在重复测量设计时对二元逻辑回归进行计算。
15. 模型设定。模型设定应是通过关于内容的统计学标准，而不是通过关于形式的算法来推导的。各个预测变量应相互不相关。很多作者明确建议不要使用自动变量选择的方法，但是在有保留的情况下，作为一种探索性方法也是可以使用的。对于这两种处理

方法，建议做如下处理：首先让内容上的有关预测变量进入模型，然后通过显著性检验剔除统计学上的无关变量。如果模型含有预测变量之间显著的交互作用，则预测达不到显著性的预测变量也保留在模型中。

16. 例如，逐步法是根据形式上的标准（统计学上的关联）进行工作的，不适用于理论推导的建模，因为逐步法也选择了内容上没有关联的预测变量。应根据可信的、关于内容的标准，对纯粹探索性的或者预测性的工作方法进行交互检验。向后法应优先于向前法，因为向后法与向前法相反，是从检验一阶交互作用开始的，因此不存在仓促剔除潜在的抑制变量的风险。但是，逐步法不消除多重共线性，因此至少要通过交叉验证予以保障。
17. 分界值（分割点值，分类阈值）。分界值主要用于减小预计发生的总风险，应与具体问题协调一致。SPSS 预设的分界值为 0.5，表示错误分类具有同样的成本，也就是假阳性或者假阴性，只有在某些特定情况下才能不经修改地予以采用。不同大小的分界值造成不同的分类和分类表。分界值的大小由其功能决定，例如，敏感度和特异度之间的权衡。较高的分界值（ >0.5 ）减小敏感度并增大特异度；较低的分界值（ <0.5 ）增大敏感度并减小特异度。如果误判的成本同样大，则分界值等于 0.5。通常，假阴性的成本超过假阳性的成本。通过给误判（假阳性或者假阴性）分配不同的成本（亏损），可以更精确地测定分界值。向具有错误分类（假阳性或者假阴性）的单元格分配了各自的成本或者成本比例。在根据变化的分界值多次实施这个过程时，可以将错误分类的成本和数量相乘所得的各个乘积加起来，从而测定出哪个模型的额外成本最小。
18. 离群值和高杠杆值。如果测定的回归方程对数据的拟合不佳，则可能发生个别个案实际上属于某个组，但是从输出结果来看却属于另一个组的现象（从分类图中也可以很容易识别出来）。将离群值或高杠杆值在工作文件中另存为残差或其他数值，就可以通过偏差分析识别出这些离群值或者高杠杆值（Hosmer & Lemeshow, 2000, 167-186）。对于学生化残差，应仔细调查绝对值超过 2 的；如果绝对值超过 3，则明确表明存在离群值或高杠杆值。

离群值列表（示例）：

皮尔逊残差

```
if abs (SRE_1) >= 2 AUSREISR=2.
exe.
if abs (SRE_1) >= 3 AUSREISR=3.
exe.

temp.
select if AUSREISR >= 2.
list variables= ID SRE_1.
```

形式上的显著数值不一定是内容上的显著数值（对于社会科学数据尤其如此）。

SPSS 将学生化残差（也就是皮尔逊残差）标为“标准残差”（变量 SRESID），而将标准化残差标为“归一化残差”（变量 ZRESID），这就有可能在一定程度上产生混淆。此外，SPSS 还可以保存反映出个案对预测值影响的统计量，主要是 Cook、杠杆值和 DfBeta。下面这个例子就获取了 Cook 统计量。

离群值列表（示例）：

Cook 统计量

```
if COO_1 >= 1 AUSREISC=1.
exe.
if COO_1 >= 2 AUSREISC=2.
exe.

temp.
select if AUSREISC >= 1.
list variables= ID COO_1.
```

通过直方图或者箱图来表现这些参数，可以方便地识别出离群值。例如，本例获取了皮尔逊残差的正态分布。

正态分布（示例）：

```
EXAMINE
  VARIABLES=SRE_1
  /PLOT BOXPLOT HISTOGRAM NPLOT
  /COMPARE GROUP
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

下面的处理方法可以识别出成组离群值或者显著数值。通过皮尔逊残差的平方测定 Δ 卡方，后者可以对离差的变化做出估计。

Δ 卡方

```
compute DELTACHI=sre_1*sre_1 .
exe.
GRAPH
  /SCATTERPLOT (BIVAR) =
    PRE_1 with DELTACHI BY ID (name) .
```

如果用预测值截取 Δ 卡方，则离群值可以对模型拟合做出估计。这两条曲线展现了变量类别分别为 0 和 1 的因变量的两个类别。尤其是对于预测概率较高时的 0 曲线或者预测概率较低时的 1 曲线，偏差值越高，则模型拟合越差。

$\Delta\beta$

```
GRAPH
  /SCATTERPLOT (BIVAR) =
    PRE_1 with COO_1 BY ID (name) .
```

通过用预测值截取 Cook 统计量（也称作 $\Delta\beta$ ），也可以类似地对模型拟合做出估计。插图大致反映出了 Δ 卡方。但是，在点云外部的高杠杆离群值十分引人注目，应将其从分析中剔除。

预测 vs 残差

```
GRAPH
  /SCATTERPLOT (BIVAR) =
    PRE_1 with sre_1 BY ID (name) .
```

类似地，通过截取预测概率和标准残差，可以很方便地识别出每个组中的离群值。

19. 模型拟合优度（错误分类）。通过最佳的估计，模型应可以正确地重现大部分观察到的事件。但是如果测定的回归方程对数据的拟合不佳，则可能发生个别个案实际上属于某个组，但是从输出结果来看却属于另一个组的现象。例如，在分类图中（也可参见下文关于离群值的论述）。每次在判断个案归属于哪个组时，数值低于 80% 是不可接受的，根据应用领域不同，甚至可以提出更高的要求。应检验错误分类的个案和成本（假阴性的成本通常高于假阳性的成本），必要时相应地调节分类阶段。所观察的命中率是否超过随机水平，可以利用二项检验进行检查（参见 Menard, 2001, 34）。
20. 检验预测优度（排除过度拟合）。一个模型在完成参数化之后，应检验其预测模型的优度是否具有实际关联性。尤其是排除了模型将误差数量增多的可能性。除了分类图（见上文）之外，也可以进行交叉验证。如果利用创建模型所基于的样本（称为“训练数据”）对模型进行了检验，则命中率可能估计过高（过度拟合），尤其是在特殊的模型中会出现过度拟合现象。其原因通常是训练数据集（乖离率，分布等）的特殊性。因此，应根据验证数据，始终通过交叉验证来检验模型中是否存在过度拟合。交叉验证是利用一个或者多个其他（子）样本（称为“验证数据”）对模型进行的检验，在 LOGISTIC REGRESSION 中可以很方便地通过 /SELECT 选项调用[变量和条件]。如果一个模型表现出很大的性能差异，例如，用训练数据可以将 80% 的数据正确分级，但是利用验证数据时这个比例可能只能达到 50%，则就存在过度拟合。相反的现象就称为拟合不足，也就是忽视了真实的数据现象。拟合不足现象主要发生在过于简单的模型中。
21. 在解释回归系数和胜率（Exp(B)）时的特殊性。（a）预测变量的尺度水平：对于胜率和回归系数的解释，其区别在于分类预测变量和定量预测变量。对于定量变量，可以用整个统一定义域的一个公共值来表达其影响；对于分类变量，则测定 $n-1$ 个变量类别或单位的数值。需要注意的是，编码会对胜率或回归系数的大小，或正负号起作用（例如，二元因变量的系数可能正负号颠倒，对此参见关于参考类别的注释）。（b）分类预测变量的编码：预测变量的编码对回归系数的解释及其计算有影响。如果对于病例个案或者事件个案，编码偏离 1，并且对于对照个案，编码偏离 0（对此参见关于参考类别的备注），则必须用另外的方法测定参数。（c）非标准化回归系数对比标准化回归系数。非标准化回归系数可能与标准化回归系数有很大区别，并且完全错误地反映了各自预测变量的影响。Menard（2001）建议对于分类变量和带有自然单位的变量采用非标准化回归系数或胜率，对于没有共同单位的定比数据采用标准化回归系数。通过在分析之前将预测变量本身标准化，就可以针对定量预测变量的模型提取出标准化回归系数。然后就可以将提取出的回归系数解释为标准化回归系数。对于带有分类预测变量的模型，处理起来就更为复杂（参见 Menard, 2001）。
22. 在线性回归中，通常建议将标准化回归系数用于比较在一个样本/总体内部的定量变量，或者用于没有共同单位的定量变量。对于后者应考虑到，其测定可能是取决于所选择的样本，并且根据模型拟合优度不同，只能有所保留地将这种测定结果普遍化。非标准化回归系数建议用于比较样本/总体之间的定量变量，或者用于具有自然/共同单位的定量变量。根据这两种回归系数的优点和缺点，Pedhazur（1982², 247-251）建议给定两种量度。如果在分析之前将数据 z 标准化，则将 β 值给定为 B 值。

23. 类别水平的完整。如果有数据缺失，结果中部分在括号里给出的变量类别与编码不一致。例如，如果数据被编码为 0 至 7，但是只有变量类别 0、1、5 和 7 存在，则不显示编码（0）、（1）和（5），而是显示（1）、（2）和（3）（最高的编码是冗余的）。因此，建议仔细检验类别水平是否完整，以确定所测定的参数与正确的类别类别相联系。例如，在这个例子中，给出的变量类别（1，第一级）可能与编码（1，第二级）相混淆。为了辨别清楚，应不断查看随之输出的表格“分类变量的编码”。
24. 协变量模式的数量。具有多个变量类别的预测变量越多，则协变量模式以及由此所需的个案数量就越大。逻辑回归的出发点是，所有的单元格都已占满。空白的或者未占满的单元格会导致在解释基于卡方的统计量时出现问题。如果个案数量超过协变量模式的数量，则应始终成组地测定模型参数。例如，通过 NOMREG 中的皮尔逊或者偏差。因此，作为拟合优度量度的方差或者皮尔逊，不适合带有定比预测变量的模型。如果协方差模式的数量大致等于个案的数量，则应始终逐个案地测定模型参数（例如，通过成组的 Hosmer-Lemeshow 检验，只有在 SPSS 过程命令 LOGISTIC REGRESSION 中可以使用，为此，多项模型必须分解为二元模型）。对于个案数据，Hosmer-Lemeshow 检验应始终优先于离差或者皮尔逊这两个拟合优度量度。

3.3 有序回归

有序回归提出的问题是针对根据等级排列（“较大-较小”）的相依事件发生的概率（Hosmer & Lemeshow, 2000, 288-330）。根据等级排列的相依事件的概率可能是分类的连续变量或者定序评估，举例如下。

- 根据哪些实验室参数，可以将癌症患者归类到 0 至 4 级？
- 如果将消费者对一种新上市产品可能给出的反应分为“完全没有”、“几乎不”、“平淡”、“良好”和“强烈”，是否可以辨别出消费者相应的特征？
- 由于哪些产品特征，决定某种产品的销售情况“很好”、“良好”、“差”或者“非常差”？
- 根据哪些参数，可以将一个刚刚失业的人归类到“短期失业者”（0~6 个月）、“中期失业者”（7~12 个月）或“长期失业者”（>12 个月）？
- 疗法的哪些特点对于治愈皮肤组织烧伤具有最佳的效果。例如，在 0~10cm²、11~20cm²、21~30 cm² 和 31~40cm² 面积的皮肤上测量到的？

如果因变量是定序的，则根据尺度水平的解释不同可以使用多种分析方法（参见 Menard, 2001、Tutz, 2000）。

- 评估为定类数据：忽略了类别水平的信息。与此相关的信息丢失问题是，可能估计出的参数数量比原本所需的更多，这又增大了非显著性事件的风险。适合的分析方法主要有多元回归和判别分析（参见 Klecka, 1980、Press & Wilson, 1978）。
- 评估为定序数据：观察到的定序数据是基础（至少是）定距数据的一种表现（变体 I，分类连续变量）。因变量是真正的定序变量，其顺序不代表大致的类别（变体 II，根据等级排列的评估）。适合的分析方法主要是有序回归或者 LISREL（WLS）。通常

不建议将定序数据分析为定距型，因为超出允许范围的信息富余会从量度理论上导致解释出现问题。下面介绍有序回归方法。

3.3.1 有序回归方法和与其他方法的比较

SPSS 提供的有序回归是建立在阈值模型（又称比例优势模型，关于其他变体参见 Hosmer & Lemeshow 的著作，2000，第 8.2 节）。这个模型的基本假设是，可观察的有序因变量（显变量）是一个实际上定距潜变量的分类表现，通过解释变量的线性组合可以测定其期望值。显变量和潜变量之间的联系是基于阈值模型。在这里首先假设，连续潜变量是基于所调查的过程，但是只能观察到基础变量的分类并且大致分级的表现。根据观察到的因变量分布（基础：累积概率模型）选择了一个连接函数（又称链接函数，也称为“identity function”），以便能够通过适当地描述自变量和因变量之间的关系而测定出一个连续体。

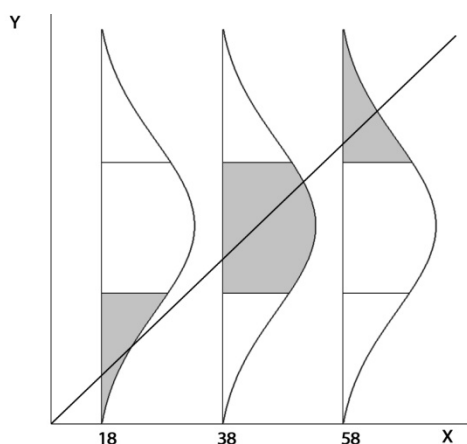
对于逻辑回归（二项、多项）模型不需要做这个测定，因为根据推测，Logit 分布的回归方程是线性的。对于有序回归，必须检验因变量各个类别的发生概率，并且在实施有序回归之前确定这个概率。例如，如果各个类别是平均分布（概率）的，则选择 Logit 函数；如果相反较高的类别具有较大的概率，则选择互补重对数函数等。对于有序回归的各种阈值，根据 Tutz（2000）的著作选择累积的 Logit 模型、Probit 模型和极值模型，这些模型都是 SPSS 可以提供的。

潜在连续体在这种情况下被分成分段；此时“阈值”决定了哪个潜在值（曲线位置）能够归属于因变量的哪个可观察类别。很容易理解的是，阈值的数量始终比类别的数量少一个：一个阈值创建两个类别，两个阈值创建三个类别，三个阈值创建四个类别等。因此，通过用这些阈值进行调整，就能够判断连续体上的各个个体位置应分配给哪个类别。如果连续体上的一个个体位置低于第一个阈值（直线），则将其分配给第一个类别；如果在潜在连续体上有一个个体位置超过第二个和第三个阈值，则将其分配给第三类别。这样的连续体就可以想象成一条密度曲线，由 $k-1$ 个表现为垂直线的阈值分割成 k 个类别。曲线下方的面积相对于各个类别的（累积的）事件概率（总共 1.0 或者 100%，通过连接函数测定）。在各个模型内部，各个类别（曲线部分面积）代表了单个预测变量等级（单因子模型）或者预测变量等级组合（多因子模型）的事件发生概率。因此，对于每个个案和预测变量等级（单因子模型）或者预测变量等级组合（多因子模型）而言，在这些前提条件下将事件分配给因变量其中一个类别的概率始终为 1.0（100%）。如果通过每个个案的链接函数测定了各个类别的概率，则在累积概率的基础上测定出概率最大的类别，也就是针对模型中所包含各个预测变量等级的任意组合。因此，对于每一级的因变量，同样也有累积概率。第一级基于第一个累积概率，第二级基于第二个累积概率，直到最后一个，最后一个类别始终含有累积概率 1.0（对此的解释参见例 2）。

示例

回到 Tutz 著作（2000，第 6 章）中关于失业（“y”，“短期失业”、“中期失业”和“长期失业”）的类似例子，这个例子调查了变量“失业者年龄”（定距“x”）对具有三个变量类别的“失业持续时间”的影响（对此参见 Hosmer & Lemeshow 的著作，2000 年，第 298-303 页关于母亲体重和新生儿体重等级的例子）。关于失业持续时间的例子从两方面做出假设：可观察的定序变量“失业持续时间”恰当地表现了定距型构件，只是为了解释失业者随着年龄的增长实际失业时间越来越长。例如，对于老年人的“长期失业”类别，可以提出针对

可信性的某些反对意见。这些假设共同构成了一个假设，这个假设既是关于根据年龄可以预计某人归属于哪个类别，又是关于相应的概率，也就是具有“短期失业”、“中期失业”和“长期失业”三个类别的一个期望模型。



在 X 轴上，并排地截取失业者越来越大的年龄值。在这些年龄值上方，各自垂直地截取一条具有同样类别的潜变量的期望值密度曲线（在图中只在年龄值 18、38 和 58 上方）。现在如果除了并排的密度曲线之外，再将每个年龄的潜变量期望值截取为回归直线，则这条直线就在曲线中与不同类别相交叉。不同年龄的失业者分配给表示失业可能持续时间的不同类别。如果用分类显变量的定序来表达，就是随着年龄的增大，被划分到最低的定序等级“短期失业”的概率逐渐减小，而被划分到最高的定序等级“长期失业”的概率逐渐增大。如果用显变量的定距来表达，就表示随着年龄的增大，长期失业的期望值逐渐增大。

Tutz (2000) 确定了有序回归与名义模型（如多项逻辑回归）和定量模型（如线性回归）的界限。根据他的著作，与分类名义模型（如多元回归）相比，参数可以需求量更小、更简单地对于有序回归做出解释；此外，在利用分类名义模型分析有序信息时，会丢失关于类别次序的重要信息。从这一点来看，有序回归也可以理解为多项逻辑回归的一种扩展形式。如果用定量模型（如线性回归）对（在某些情况下也是多次分级的）有序信息进行分析，相反就有生成一个臆造结构的风险。

以高要求的定序型类别进行有序建模，可避免其他的保守假设，尤其是正态分布和误差分布。在与时间相关的分析中，有序回归由于具有更好的适用性，甚至超过了简单线性模型。利用有序回归可以解决提出的下列问题：预测一个事件或者分类，检验预测变量的关联性（必要时还要检验预测变量相互之间的交互作用或者预测变量与协变量的交互作用），检验一个分类模型的拟合优度，估计预测变量。

下面首先是带有两个定距预测变量的一个例子，然后是带有两个分类预测变量的一个例子。

3.3.2 例1 界面操作和语法：定距预测变量（WITH-选项）

关于例子的简要说明

对于这些患者，除了身体质量指数（变量“BMI”，定序）之外，还要采集关于体重（变量“KGEWICHT”，二元）、生育孩子（变量“EKINDER”，二元）、臀围（HUEFTE，区

间)和腰围(TAILLE, 区间)的数据。

例 1 调查了两个定距变量(腰围和臀围, 单位: cm)对身体质量指数的定序量度的影响。变量“BMI”分为四个类别: “ ≤ 20 ”、“20-25”、“26-30”和“ > 30 ”。

例 2 调查了两个分类变量(KGEWICHT: 体重, 单位: kg、EKINDER: 生育孩子(是/否))对身体质量指数(变量“BMI”)的定序量度的影响。EKINDER 是定类变量。KGEWICHT 分为“39-66”和“67-120”两个类别, 严格来说是一种定序变量。

例 1: 定距预测变量

例 1 调查了两个定距变量(腰围和臀围, 单位: cm)对妇女身体质量指数的定序量度的影响。变量“BMI”分为四个类别: “ ≤ 20 ”、“20-25”、“26-30”和“ > 30 ”。为了能够选出一个合适的链接函数, 首先利用一个简单的条形图显示出因变量的分布; 然后, 利用肯德尔等级相关系数验证在至少三个定序变量之间是否存在实质性相关; 最后, 再对有序回归进行计算。

界面操作(示例)

在调用有序回归之前, 通过“交叉表”和肯德尔量度等级相关系数来检验变量之间是否存在实质性关联, 从而确定连接方式(链接函数)。通过一个条形图可以看清楚因变量的分布。作为结果, 针对有序回归选择连接方式“Probit”。

在 SPSS 程序主界面选择以下菜单项: 分析 → 回归 → 有序

把变量“BMI”拖入窗口“因变量”, 把变量“huefle(臀围)”和“taille(腰围)”拖入窗口“协变量”。



子窗口“选项”。首先在“标准”一项下接受所有的预设置, 例如, “最大迭代次数”(100)、“最大对分步数”(5)、“对数似然收敛性”(0)、“参数收敛”(0.000001)、“收敛区间”(95)、“Delta”(0)和“奇异性容许值”(0.00000001)。在“连接方式”一项下选择“Probit”。单击“继续”按钮。

子窗口“输出结果”。选定“拟合优度统计量”、“摘要统计量”、“参数估计值”和“单元格信息”。

调用预测的和实际的类别概率作为“保存的变量”。单击“继续”按钮。

子窗口“类别”。将“主效应”确定为模型。“F”指的是因子, “C”指的是协变量。

单击“继续”按钮。

单击“确定”按钮开始计算。

语句:

```
NONPAR CORR
  /VARIABLES=bmi huefte taille
  /PRINT=KENDALL TWOTAIL NOSIG
  /MISSING=PAIRWISE .

GRAPH
  /BAR (SIMPLE) =pct BY bmi .

PLUM
  bmi WITH huefte taille
  /CRITERIA = CIN (95) DELTA (0) LCONVERGE (0) MXITER (100)
  MXSTEP (5) PCONVERGE (1.0E-6) SINGULAR (1.0E-8)
  /LINK = PROBIT
  /PRINT = CELLINFO FIT PARAMETER SUMMARY
  /SAVE = PCPROB ACPROB .
```

备注: 用带有 PRINT=KENDALL 选项的 NONPAR CORR 命令调用肯德尔相关系数相关量度; 进一步细节可参见关于非参数相关量度的一节 (Schendera, 2004)。

GRAPH 命令利用 /BAR (SIMPLE) 子命令调用因变量 BMI 的一个条形图。

PLUM 命令调用有序回归方法。PLUM 命令使定序因变量 BMI 进入模型, WITH 命令将两个定距预测变量 HUEFTE (臀围) 和 TAILLE (腰围) 纳入模型。模型 “BMI with HUEFTE TAILLE” 只调查主效应, 不调查预测变量之间的交互作用。

在 CRITERIA 一项下, 主要是对估计算法进行调整。利用 CIN (95, 预设置) 定义置信区间为 95%; 在 CIN 一项下可以给定从 50 到 99.99 的数值。在 DELTA 一项下可以给定 0 到 1 之间的数值, 然后将这个数值添加到空白单元格, 从而确保了估计算法的稳定性。在本例中, 空白单元格不应加入数值。DELTA 不应与 BIAS (不使用) 混淆。在 BIAS 一项下可以类似地给定一个在 0 到 1 之间的数值, 然后将这个数值添加给所有观察到的单元格频率。MXITER (100) 和 MXSTEP (5) 给定了迭代最大次数和最大对分步数 (这些数值必须是正整数)。在达到设定的最大次数后, 估计迭代停止。通过 LCONVERGE 和 PCONVERGE, 将用于对数似然值函数和参数估计的收敛标准传递给 SPSS。在 LCONVERGE 一项下, 给出达到对数似然值收敛性的一个阈值, 除了在等于 0 时, 通常这个数值采用科学计数法: 1.0E-1、1.0E-2、1.0E-3、1.0E-4 和 1.0E-5 (预设置为 0)。当最后两次迭代之间的对数似然值绝对或相对变化小于这个值时, 迭代过程结束。在预设定为 0 时, 不使用这个标准。在 PCONVERGE 一项下, 给出达到参数收敛性的一个阈值, 除了在等于 0 时, 通常这个数值采用科学计数法: 1.0E-4、1.0E-5、1.0E-6、1.0E-7 和 1.0E-8 (预设置为 1.0E-6)。如果参数估计值中的绝对或者相对变化小于这个值, 则可以认为这个算法达到了正确的估计值。在 PCONVERGE (0) 时, 不使用这个标准。在 SINGULAR 一项下可以给定奇异性检验的容许误差; 这个数值采用科学计数法: 1.0E-5、1.0E-6、1.0E-7、1.0E-8、1.0E-9 和 1.0E-10 (预设置为 1.0E-8)。

在 **LINK** 一项下，可以确定用于转换累积概率的链接函数，以便对有序回归模型做出估计。根据探索的结果，选择了 **Probit** 函数。除了预设置的 **Logit** 函数（**LOGIT**，典型用途：均匀分布的类别）之外，还可以选择 **CAUCHIT**（逆柯西函数，典型用途：潜变量具有很多极端值）、**CLOGLOG**（互补重对数函数，又称为 **Gumbel**，典型用途：较高的类别具有较大的概率）、**NLOGLOG**（负重对数函数，典型用途：较低类别具有较大的概率）和 **PROBIT**（**Probit** 函数，典型用途：潜变量是正态分布的）。如果经验分布只在很小程度上符合提供的模型函数，则可以将因变量的各个类别合并起来，只要从概念上允许，就可以使其符合预设的模型函数中的一个。

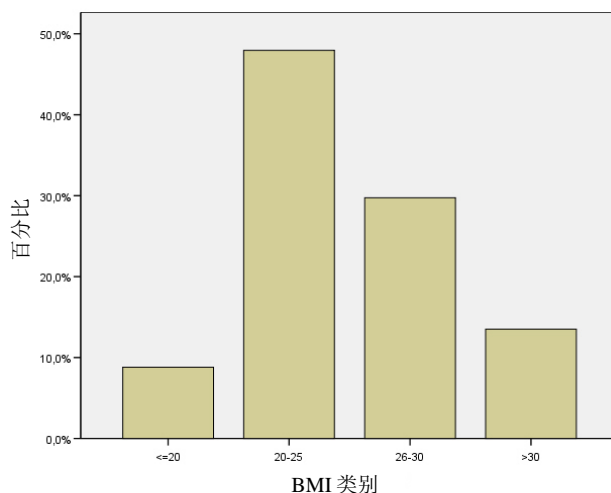
在 **PRINT** 一项下设置 **SPSS** 输出结果。在本例中，**CELLINFO** 调用单元格信息（尤其是观察频率和预期频率，定距预测变量或者具有多个变量类别的因子），**FIT** 调用两个拟合优度卡方统计量（皮尔逊，似然比），**PARAMETER** 调用参数估计值、标准误差、显著性和置信区间，**SUMMARY** 调用模型总结（**Cox & Snell**、**Nagelkerke** 或 **McFadden** 拟 R^2 ）。例 2 介绍了选项 **TPARALLEL**。其他选项（在本例中没有使用）包括 **CORB** 和 **COVB**（参数估计值的渐近式相关系数矩阵或协方差矩阵）、**HISTORY**（迭代过程的表格式记录）和 **KERNEL**（代替完整的对数似然值函数，只调用对数似然值函数的核心，也就是没有多项常数）。

通过 **SAVE** 可以将计算出的估计统计量保存到当前数据集，以便利用 **PCPROB** 在预测类别中把估计概率，即一个因子形式分类，利用 **ACPROB** 在实际类别中把估计概率，即一个因子形式分类。例 2 介绍了其他选项。

通过额外的子命令和选项（例如，**LOCATION**、**SCALE** 和 **TEST**），**SPSS** 过程命令 **PLUM** 可以将这里所介绍的例子改进成借助于语句的有序回归。对此请参见 **SPSS** 语法文档和统计学专业文献。

3.3.3 输出结果和解释

图形



可以（首先）对条形图做出假设，即定序因变量 **BMI** 的变量类别大致呈正态分布。从输出的条形图可知对于定距型（未分组）潜在 **BMI** 值接近呈正态分布的假设是不实际的，因此，

选择 Probit 函数作为链接函数。如果从概念上允许将因变量的两个最低类别合并起来，则很明显更符合模型函数 NLOGLOG，并且在某些情况下会得出更好的离散值。

非参数相关性

相关性			BMI 类别	臀围 (cm)	腰围 (cm)
肯德尔等级 相关 tau-b 系 数	BMI 类别	相关系数	1.000	0.628**	0.615**
		显著性 (两侧)	.	0.000	0.000
		N	659	634	634
	臀围 (cm)	相关系数	0.628**	1.000	0.605**
		显著性 (两侧)	0.000	.	0.000
		N	0.634	637	637
	腰围 (cm)	相关系数	0.615**	0.605**	1.000
		显著性 (两侧)	0.000	0.000	.
		N	634	637	637

** 在 0.01 水平上，相关性是显著的 (两侧) 。

从表 “相关性” 可以看出，在变量之间存在显著性相关，没有什么因素与真正的有序回归的计算相冲突。在根据位置估计值解释因子的预测方向，相关分析是很有用的。

PLUM——有序回归

针对真正的有序回归的 SPSS 输出结果，首先从关于零频数的警告开始。原因是，模型含有定距预测变量，并且某些拟合统计量试图将因变量的观察值分配给预测变量 (HUEFTE 臀围, TAILLE 腰围) 的每个组合。但是，由于这些预测变量的尺度十分精细，由此产生很多种组合可能性，因此很多单元格组合之后还是保持空白状态。所以，从因变量的角度来看，观察次数太小了；从预测变量的角度来看，其组合 (也就是协方差模式) 的数量太大了。因此应删除几个预测变量，或者将因变量或自变量的几个类别合并起来。或者如果有可能的话，将更多个案输入模型。

警告

有 1283 个 (71.8%) 具有零频数的单元格 (即所决定的因变量的水平超过预测变量数值的组合) 。

务必要考虑到这条说明，因为对于很多带有零频数的单元格而言，模型拟合的有效性是不清楚的，并且很难解释基于卡方的拟合统计量 (如对数似然值)。由于上述原因，关于单元格的表格 (通过 CELLINFO 选项调用) 对于例 1 来说范围太大，因此在这里没有反映出来 (对此参见例 2) 。

个案处理总结

		数量	边际百分比
BMI 类别	<=20	58	9.1%
	20-25	302	47.6%
	26-30	189	29.8%
	>30	85	13.4%
	有效	634	100.0%
	缺失	29	
	总值	663	

表“个案处理总结”给出了模型的分类变量，例如，在这里除了因变量 BMI 的绝对频数和百分比频数之外，还有缺失值（N=29）。

模型拟合信息

模型	2*对数似然值	卡方	自由度	显著性
只有常数项	1417.108			
最终	827.371	589.737	2	0.000

链接函数：Probit。

表“模型拟合信息”表明了模型是否有能力做出精确的预测。“卡方”值表明，带有预测变量的模型（包括预测变量 HUEFTE 臀围和 TAILLE 腰围的总模型“最终”）是否能比没有预测变量的模型，也就是常数模型（“只有常数项”）提供更好的信息。

针对常数模型（“只有常数项”）和总模型（“最终”）输出 2*对数似然值。如果在模型中出现很多空白单元格（零频数），则对数似然值数据本身可能就是值得怀疑的；但是，对数似然值之间的差值通常总是解释为（接近于）卡方分布的（McCullagh & Nelder, 1989）。所给出的卡方（589.737）是基于“只有常数项”模型（常数模型，1417.108）和“最终”模型（总模型，827.371）的对数似然值之间的差值。卡方的显著性（p=0.000）表明，带有预测变量的模型能够比纯粹的常数模型提供更好的信息。卡方显著性是一种符合期望的显著性。下面这个关于模型拟合优度的表格介绍了总模型在多大程度上得到了改善。

拟合优度

	卡方	自由度	显著性
皮尔逊	1315.705	1336	.649
离差	739.990	1336	1.000

链接函数：Probit。

表“拟合优度”表明，实际的、观察的单元格频率是否和在多大程度上与利用模型计算出的期望概率有显著区别。卡方检验（皮尔逊，离差）的 p 值为 p=0.649 或 1.000，都不显著，这就表明计算出的模型具有很高的拟合优度。在这些检验时，不希望达到显著性，因为这表明模型与数据有统计学上的重大偏差。根据 Menard（2001）的著作，对于逻辑回归而言，离差量度比皮尔逊量度的信息量更大。但是，由于离差量度不仅有对自变量的辨别能力，而且能对基本比率（例如，可能不对称地占据因变量类别的单元格）做出反应，也要考虑使用

McFadden R^2 或者似然比检验。与偏差相反,这两个检验都是基于常数模型和最终模型之间的对数似然值比值;不对称的分组大小不会对检验统计量起作用。当占据的单元格很少或者甚至空白的单元格很多时,使用卡方检验是有问题的(可参见警告提示,这也适用于 McFadden R^2)。离差的卡方值和自由度的卡方值的商表明存在超聚集性($739.99/1336=0.55$)。

拟 R^2	
Cox & Snell	0.606
Nagelkerke	0.666
McFadden	0.387

链接函数: Probit。

从关于拟 R^2 统计量的表格可以查出,通过模型解释的方差占多大比例。目标值大约为 1.0 或 100%。拟 R^2 基本上等于定量回归的 R^2 。Cox & Snell、Nagelkerke 和 McFadden 拟 R^2 是近似值(详情参见下文)。根据 Menard 的著作(2001, 2000), McFadden 拟 R^2 对于逻辑回归是最适合的量度。McFadden R^2 独立于基本比率和样本量,从概念上最接近于最小二乘 R^2 ,适合多元的定类因变量、有序因变量,也适合二元因变量,从而也适用于对模型的比较(前提条件是满足了卡方方法的假设)。McFadden R^2 表达了对数似然值的呈比例缩减,在有关文献中,在 0.2 到 0.4 之间的这些值被视为是非常令人满意的。得出的 R^2 (0.387) 表明模型拟合比较好,从而表明自变量完全可以解释出因变量归属于哪个组。Nagelkerke R^2 表明模型解释了大约 67% 的方差。

各个量度之间的最主要差异是基本数据、理论最大值和解释。作为 Cox & Snell R^2 的校正形式, McFadden R^2 和 Nagelkerke R^2 的数值为从 0 到 1。相反, Cox & Snell R^2 的数值在一个完美的模型中达不到 1。这两个量度都是基于似然值,可以解释为方差解释,但是也分别受到样本量和基本比率的预测。与这两个量度相比, McFadden R^2 是基于预测变量模型的卡方值除以常数模型的对数似然值(Log-Likelihoods, -2LL)所得出的商,在 0 到 1 之间变化,不受样本量和基本比率(具有或者不具有所调查特征的个案所占比例)的影响。McFadden R^2 可以解释为相联的规模,或者与 PRE 量度类似,可以解释为错误率降低的百分比。1 表示完美的预测或者关联, 0 表示没有预测力或者不存在关联。

McFadden R^2 说明

在有关的统计学书籍以及 SPSS 书籍中,对 McFadden R^2 的阐述并不一致。诺贝尔经济学奖得主 Daniel McFadden 教授(美国伯克利大学, Pers. Kommunikation 2004.01.27)很高兴对他发明的 R^2 做一番简短的论述。

定义。McFadden R^2 在 0 到 1 之间取值。在特殊情况下, McFadden R^2 可能是负数。McFadden R^2 的定义是 $= 1 - LL(ML) / LL$ (“比值”)。LL 代表对数似然值。MLE 表示利用 MLE (最大似然估计法)的估计。“比值”表示利用常数比值的估计。利用“比值”进行估计,一次观察的对数似然值就是样本比值乘以样本比值对数所得乘积的备选方案之和。这相当于对只含有备选方案特定常数的模型进行一次最大似然估计。

如果完整的模型函数含有的与线性无关的备选方案特定常数数量最多,则 McFadden R^2 在 1 和 0 之间。如果所有斜率参数不是显著的,则 McFadden R^2 接近 0。如果模型近似做出完美

的预测，则 McFadden R^2 接近 1。如果在整个模型中没有整组的备选方案特定常数，则 McFadden R^2 也可以是负数。在任何情况下，McFadden R^2 都不能大于 1。McFadden R^2 的这个特性与一般 R^2 相同。除了似然比的渐近分布之外，McFadden R^2 与卡方不存在任何关联。

解释。McFadden R^2 是“拟合优度”的一项指标，与传统 R^2 类似，也就是说，表示不取决于尺度和样本地对模型拟合优度的一个常规检验统计量进行转换，在这种情况下，是对似然比检验的转换。和传统 R^2 一样，McFadden R^2 作为一个量度是很有帮助的，可以评估模型拟合或者具有同样因变量的替代模型的相联。但是，如果想要将带有不同因变量的模型相互比较，则 McFadden R^2 可能有很大的误导作用。

如果想象带有一个常数项和一个斜率系数的二元模型，以及在备选方案之间平均分割（0.5）的观察值，则可以想象出 McFadden R^2 尺度。如果以 0.8、0.9 或者 0.95 的概率预测观察到的备选方案，则 McFadden R^2 的相应数值为 0.28、0.53 和 0.71。但是，如果观察值分为 0.7 比 0.3，则同样的预测概率（0.8、0.9 和 0.95）就会得出 McFadden R^2 值为 0.18、0.47 和 0.68。

补充说明结束

McFadden R^2 通常被视为比较适合用于对模型的比较（详情参见上文），否则就如同在这个个案中一样，可能无法简便地对这个参数做出解释。所实施有序回归的主要结果参见“参数估计值”表。

参数估计值

		估计值	标准误差	Wald 统计量	自由度	显著性	置信区间 95%	
							下限	上限
阈值	[bmi = 1.00]	11.897	0.746	254.505	1	0.000	10.436	13.359
	[bmi = 2.00]	14.510	0.815	316.737	1	0.000	12.912	16.108
	[bmi = 3.00]	16.323	0.869	352.744	1	0.000	14.619	18.026
位置	huefte	0.073	0.007	106.124	1	0.000	0.059	0.087
	taille	0.079	0.009	77.509	1	0.000	0.062	0.097

链接函数：Probit。

表格参数估计值针对计算出的模型总结了估计值、标准误差、Wald 统计量、自由度、显著性和置信区间。对于因变量 BMI 的四个类别，首先列出了三个阈值及其估计值。阈值估计值应做如下解释：如果一个数值低于第一个阈值 11.897，则其归入类别 1；如果一个数值在 11.897 和 14.510 之间，则其归入类别 2，依此类推（关于相应置信区间的意义参见下文）。两个定距预测变量 HUEFTE（臀围）和 TAILLE（腰围）的预测用这两行中的“位置”来体现。对于评估自变量的影响，变量 HUEFTE（臀围）和 TAILLE（腰围）的估计值是非常重要的（也就是位置估计值）。从“显著性”一列可以查出，哪些自变量对于模型是重要的。在本例中，HUEFTE（臀围）和 TAILLE（腰围）的显著性值分别为 $p=0.000$ ，这表明它们对因变量有统计学上的重要影响。只有正数的位置估计值（0.073 或 0.079），这表示，如果 HUEFTE（臀围）值或者 TAILLE（腰围）值增大，就会归入更高的 BMI 类别。对于负数的位置估计值

应做相反的解释，如果 HUEFTE（臀围）值或者 TAILLE（腰围）值增大，就会归入更低的 BMI 类别。

对于一位具有参数 HUEFTE（臀围）=75 和 TAILLE（腰围）=60 的妇女，可以根据线性组合：估计值 = HUEFTE*位置估计值_{HUEFTE} + TAILLE*位置估计值_{TAILLE}，然后再利用阈值估计值，从而测定出相应的 BMI 类别。例如，估计值 = (75 * 0.073) + (60 * 0.079) = 5.475 + 4.74 = 10.215。数值 10.215 低于最低阈值 11.897（并且在该阈值的置信区间之外），因此很明显归入第一个 BMI 类别。参数 HUEFTE（臀围）和 TAILLE（腰围）的数值分别为 75 和 60 的一位妇女因此归入第一个 BMI 类别。

针对 BMI 类别的阈值，还输出了置信区间。如同在本例中（BMI=1: 13.359（上限），BMI=2: 12.912（下限）所示，如果置信区间相互重叠，则明确表明在某些值域的阈值没有很好的区分作用。例如，在 12.912 到 13.359 的值域内，一个数值既有可能归入最低的，也有可能归入第二高的类别。

3.3.4 例 2 和语法：分类预测变量（BY 选项）

例 2 调查了两个分类变量（体重（单位：kg）和生育孩子（是/否））对妇女体质量标的定序量度的影响。如同在例 1 中一样，变量“BMI”分为下列四个类别：“≤20”、“20~25”、“26~30”和“>30”。变量 KGEWICHT 分为“39~66”和“67~120”两个类别，从严格意义上讲是一个预测变量。变量“生育孩子”（EKINDER）有“是”和“否”两种可能取值，是一个分类变量。通过一个简单的条形图首先显示 BMI 的分布情况，从而可以选择出适当的链接函数。利用表格统计量 Cramer V 值探索在两个因子和因变量之间是否存在实质上的关联，然后利用有序回归进行计算。

```
CROSSTABS
  /TABLES=bmi BY kgewicht ekinder
  /FORMAT= AVALUE TABLES
  /STATISTIC=CHISQ PHI
  /CELLS= COUNT .

GRAPH
  /BAR (SIMPLE) =pct BY bmi .

PLUM
  bmi BY kgewicht ekinder
  /CRITERIA = CIN (95) DELTA (0) LCONVERGE (0) MXITER (100)
  MXSTEP (5) PCONVERGE (1.0E-6) SINGULAR (1.0E-8)
  /LINK = PROBIT
  /PRINT = CELLINFO FIT PARAMETER SUMMARY TPARALLEL
  /SAVE = ESTPROB PREDCAT .
```

备注：带有选项 /STATISTIC=CHISQ PHI 的命令 CROSSTABS 主要调用 Cramer V 值，详细情况请参见关于列联表分析的章节。

GRAPH 命令利用 /BAR (SIMPLE) 子命令调用因变量 BMI 的条形图。

在前面一节，已经详细解释了 SPSS 过程命令 PLUM 的语句，在本节则只介绍分类预测变量分析的特点。根据 PLUM 命令使定序因变量 BMI 进入模型，根据 BY 命令使两个分类预测变量 KGEWICHT 和 EKINDER 进入模型。模型“BMI by KGEWICHT EKINDER”只调查了主效应，不调查预测变量之间的交互作用。

在 LINK 一项下，可以确定用于转换累积概率的链接函数，以便对有序回归模型做出估计。根据探索的结果，选择了 Probit 函数。

在 PRINT 一项下设置 SPSS 输出结果。选项 TPARALLEL 检验了平行性假设（根据这个假设，因变量所有类别的斜率是相等的），并且调用了卡方检验的平行性检验（这个选项只适用于纯粹的分类模型）。

通过 SAVE 命令，可以将计算出的估计统计量保存到当前活动数据集。例如，通过 ESTPROB 将一个因子或协变量组合归为某个类别的估计概率保存到活动数据集，通过 PREDAT 将因子或协变量组合的最大期望值对应的反应类别保存到活动数据集。

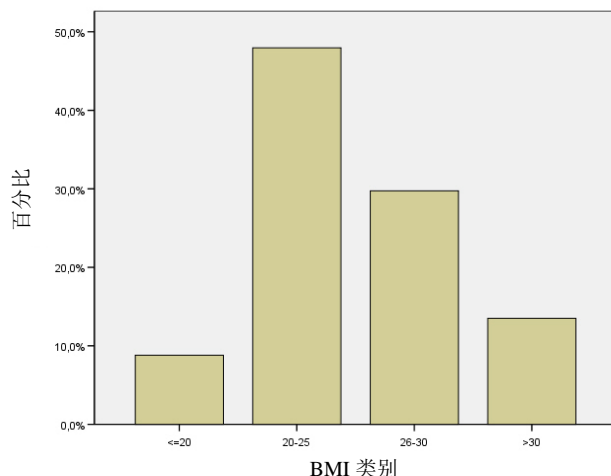
通过额外的子命令和选项（例如，LOCATION、SCALE 和 TEST），SPSS 的命令 PLUM 可以将所介绍的例子改进成有序回归。

通过子命令 LOCATION、SCALE 和 TEST 可以定义更复杂的模型。例如，如果要把交互作用或者尺度分类纳入考虑范围，则可以通过子命令 LOCATION 或者 SCALE 对其进行设定。空白的或者被删除的 LOCATION 或者 SCALE 子命令以同样的方式隐含了对简单加法模型的标准计算。自动计算出的标准模型还有常数（截距）、所有协变量（如果有的话）和呈序列排列的所有主因子。

通过子命令 TEST，可以在参数线性组合的基础上制订出呈零假设形式的假设检验。不是通过鼠标控制，而是只通过编程实现子命令 TEST 的多种功能。详细情况可以参见 SPSS 语法文档和统计学专业文献。

3.3.5 输出结果和解释

图



定序因变量 BMI 的变量类别大致上呈正态分布。对于定距型（未分组），潜在 BMI 值呈正态分布的假设是不实际的，因此，选择 Probit 函数作为链接函数。

交叉列表

在这里没有完整地表现交叉列表，而是只显示了 Cramer V 值。第一个统计量体现了体重和 BMI 之间的关联。

对称量度		数值	显著性近似值
定类尺度	Phi	0.718	0.000
定类尺度	Cramer V 值	0.718	0.000
有效个案数量		659	

- a. 没有零假设。
- b. 在假定零假设成立的情况下，使用渐近的标准误差。

第二个统计量体现了生育孩子和 BMI 之间的关联。

对称量度		数值	显著性近似值
定类尺度	Phi	0.152	0.002
定类尺度	Cramer V 值	0.152	0.002
有效个案数量		659	

- a. 没有零假设。
- b. 在假设零假设成立的情况下，使用渐近的标准误差。

从“对称量度”表中可以看出，至少在体重和 BMI 之间存在实质性相关（0.718， $p=0.000$ ），因此没有什么因素与真正的有序回归的计算相冲突。在根据位置估计值解释因子的预测方向，相关分析是很有用的。

PLUM——有序回归

针对真正的有序回归的 SPSS 输出结果首先从关于零频数的警告开始。先回想一下，模型的因变量一侧具有四个不同的变量类别，在所有分类预测变量一侧同样具有四个变量类别，因此相乘总共占满 16 个单元格。在这 16 个单元格中，有 4 个（25%）是空白的。产生四个空白单元格的原因是，对于分类预测变量的某些变量类别或者组合，不存在因变量的数值（详细情况可以从后文关于单元格信息的表格中获取）。

警告

有 4 个（25%）具有零频数的单元格（因此因变量的水平超过预测变量数值的组合）。

务必要考虑到这条说明，因为对于很多带有零频数的单元格而言，模型拟合的有效性是不清楚的，并且很难解释基于卡方的拟合统计量（如对数似然值）。

个案处理总结

		数量	边际百分比
BMI 类别	<=20	58	8.8%
	20-25	316	48.0%
	26-30	196	29.7%
	>30	89	13.5%
体重（二元）	39-66	320	48.6%
	67-120	339	51.4%
生育孩子	否	147	22.3%
	是	512	77.7%
有效		659	100.0%
缺失		4	
总值		663	

“个案处理总结”表给出了模型的分类变量。例如，在这里除了因变量 BMI 的绝对频数和百分比频数之外，还有变量的类别“体重”和“生育孩子”。

单元格信息

频数

			BMI 类别			
体重（二元）	生育孩子		≤20	20-25	26-30	>30
39-66	否	观察值	16	71	5	0
		期望值	18.943	64.571	5.342	0.144
		皮尔逊残差	-0.759	0.810	-0.153	-0.380
	是	观察值	42	167	19	0
		期望值	38.722	171.663	17.069	0.546
		皮尔逊残差	0.578	-0.716	0.486	-0.740
67-120	否	观察值	0	17	26	12
		期望值	0.070	14.283	28.291	12.355
		皮尔逊残差	-0.264	0.835	-0.618	-0.115
	是	观察值	0	61	146	77
		期望值	0.229	62.059	145.938	75.773
		皮尔逊残差	-0.479	-0.152	0.007	0.165

与例 1 相反，关于单元格信息的表格是比较清晰明了的，在这里可以表现出来。

关于单元格信息的报告总是可以从两个角度来解读：一个是关注形式的视角，一个是关注内容的视角。关注形式的视角检验了在所调查的目标事件中，是否以足够的频率出现了因变量的有关类别（尤其是在二元逻辑回归中，这个现象经常有可能导致产生令人不快的认知）。关注内容的视角检验了与对象有关的明显现象的经验分布。例如，在上面的表格中，很明显有一

定程度上的反向分布。例如，在 BMI 类别 20-25 中，体重 39-66 的妇女没有生育孩子的比例比其他所有类别都高得多，在接下来的步骤可以检验，妇女的年龄是否可能对此做出解释。

这个表格的竖列是因变量 BMI 的各个变量类别，横行是“体重”和“生育孩子”这两个因子嵌套的变量类别；除了观察频数和期望频数之外，还给出了皮尔逊残差。四个空白单元格位于最高的（即体重 39~66kg 的妇女）以及最低的 BMI 类别（即体重 67~120kg 的妇女）中，无论她们是否生育了孩子。可以认为，这些空白单元格对模型拟合的测定和基于卡方的模型拟合统计量有很明显的不利影响，此时应该考虑根据所记录的数据进行进一步的分析。

模型拟合信息

模型	2*对数似然值	卡方	自由度	显著性
只有常数项	455.019			
最终	41.634	413.385	2	0.000

链接函数：Probit。

表“模型拟合信息”表明了模型是否有能力做出精确的预测。“卡方”值表明具有两个分类预测变量的模型（“最终”）是否能比纯粹的常数模型（“只有常数项”）提供更好的信息，关于统计详细信息请参见例 1。所给出的卡方（413.385）是基于“只有常数项”和“最终”两个模型的对数似然值之间的差值，所得出的显著性（ $p=0.000$ ）表明，带有预测变量的模型能够比纯粹的常数模型提供更好的信息。卡方显著性是一种符合期望的显著性。下面这个关于模型拟合优度的表格介绍了总模型在多大程度上得到了改善。

拟合优度

	卡方	自由度	显著性
皮尔逊	3.016	7	0.884
离差	3.990	7	0.781

链接函数：Probit

从表“拟合优度”可以看出，实际观察到的单元格频率是否和在多大程度上与利用模型计算出的期望概率有显著区别。卡方检验（皮尔逊、离差）的 p 值分别为 $p=0.884$ 或 0.781 ，因此能得出不显著的结论，这说明计算出的模型具有适当的拟合优度（关于皮尔逊和离差（又称偏差）的详细情况请参见关于多项逻辑回归的一节）。在这些检验时，不希望达到显著性，因为这表明模型与数据有统计学上的重大偏差，或者有很大比例的未解释方差。但是当占满的单元格很少或者甚至有很多单元格空白时，使用卡方检验是不合适的（McCullagh & Nelder, 1989；也可参见警告提示）。

拟合优度量度表明存在超散布性或者超聚集性。卡方值离差（3.990）除以自由度（7）得出的是上一个接近于 1 的数值（0.57）。这表明存在超聚集性。

从关于拟 R^2 统计量的表格可以查出，由模型解释的方差占总体方差的多大比例。除了 Cox & Snell 之外，所有量度的目标值为 1.0 或者 100%。Nagelkerke R^2 表明模型解释了大约 51% 的方差。McFadden R^2 (0.262) 表明模型拟合比较好，从而表明自变量可以较好地解释出因变量归属于哪个组。McFadden R^2 包含一个卡方值在内，由于有较多的空白单元格，所以很难予以

解释（关于统计详细情况请参见例 1）。

参数估计值

		估值	标准误差	Wald 统计量	自由度	显著性	置信区间 95%	
							下限	上限
阈值	[BMI = 1.00]	-3.153	0.142	490.180	1	0.000	-3.432	-2.874
	[BMI = 2.00]	-0.774	0.077	101.142	1	0.000	-0.925	-0.624
	[BMI = 3.00]	0.622	0.075	69.129	1	0.000	0.476	0.769
位置	[KGEWICHT=.00]	-2.198	0.126	304.062	1	0.000	-2.445	-1.951
	[KGEWICHT=1.00]	0 ^a	.	.	0	.	.	.
	[ekinder=.00]	-0.134	0.111	1.458	1	0.227	-0.352	0.084
	[ekinder=1.00]	0 ^a	.	.	0	.	.	.

链接函数：Probit。

a. 由于这个参数是冗余的，因此设为零。

在“参数估计值”表中，有所拟合的有序回归模型的主要结果。表“参数估计值”针对计算出的模型给出了估计值、标准误差、Wald 显著性和置信区间。对于因变量 BMI 的四个类别，首先列出了三个阈值及其估计值。阈值估计值应做如下解释：如果一个数值低于第一个阈值-3.153，则其归入类别 1；如果一个数值在-3.153~-0.774 之间，则其归入类别 2，依此类推（关于相应置信区间的意义参见下文）。两个分类预测变量 KGEWICHT 和 EKINDER 的影响在“位置”行的下面，对于每个分类预测变量，这里都没有展现各自最高的类别。对于评估自变量的影响，变量 KGEWICHT 和 EKINDER 的估计值是非常重要的（也就是位置估计值）。通过观察到的预测变量类别的线性组合，例如，估计值 = KGEWICHT*位置估计值 KGEWICHT + EKINDER*位置估计值 EKINDER，位置和阈值估计值可以共同判定某个组合应归类到因变量的哪个类别。对于具有参数 KGEWICHT=0 和 EKINDER=1 的妇女，可以与例 1 类似地测定出相应的 BMI 类别。从“显著性”一列可以看出，p=0.000 的预测变量 KGEWICHT 对因变量施加了统计学上的重大影响。预测变量 EKINDER 以 p=0.227 对因变量没有施加统计学上的重大影响，因此必要时可以从模型中剔除。位置估计值是负数（-2.198 和 -0.134），这表示相关的 KGEWICHT 和 EKINDER 类别作用于较低的 BMI 类别；正数的位置估计值，则应予以相反的解释（但是不应由于 EKINDER 不显著而对其忽视）。针对因变量 BMI 的阈值，输出了置信区间。如果置信区间不重叠（如本例所示），则表示这些阈值可以清晰地区分各个类别（关于重叠类别的问题请参见例 1）。进一步的解释参见下一节。

直线的平行线检验^a

模型	2*对数似然值	卡方	自由度	显著性
零假设	41.634			
广义	37.695	3.939	4	0.414

零假设表明，关于反应类别的位置参数（斜率系数）是相互一致的。

a. 链接函数：Probit。

针对纯粹的分类模型，表“直线平行性检验”含有平行性假设的检验结果。“零假设”此时就是所实施有序回归（包括预测变量）模型的位置参数对各个类别是一类的。“常规”是一个模型，在这个模型中，不同的类别有不同的位置估计值。输出的卡方（3.939）是“零假设”和“常规”的 $2 \times$ 对数似然值之间的差值。在这个检验中，不希望出现显著性，因为这表明，位置估计值的变化范围超过了因变量的类别。输出的非显著性值（ $p=0.414$ ）表明，关于 BMI 类别的位置估计值是恒定的，无法拒绝平行性假设。

对所储存数值的逐个案解释

通过选项 ESTPROB 和 PREDAT，SPSS 将模型测定的概率保存在数据集中。ESTPROB 表明了将一个因子组合正确归类到从属类别水平的估计概率。PREDAT 表明了具有一个因子组合最高期望概率的从属等级。下表是所测定数据（通过 LIST VARIABLES=输出）的节选；对于这个例子重要的是斜体的数据行 PATID=2364 或 4795。数据不再含有概率值，而是已经含有换算出的概率。

PATID	BMI	KGEWICHT	EKINDER	EST1_1	EST2_1	EST3_1	EST4_1	PRE_1
684	4.00	1.00	1.00	.00	.21	.52	.26	3.00
2364	2.00	.00	1.00	.17	.76	.07	.01	2.00 <-
3532	2.00	1.00	1.00	.00	.21	.52	.26	3.00
3895	1.00	.00	1.00	.17	.76	.07	.01	2.00
4795	4.00	1.00	1.00	.00	.21	.52	.26	3.00 <-
5753	2.00	.00	1.00	.17	.76	.07	.01	2.00
6342	3.00	1.00	.00	.01	.25	.52	.22	3.00
7484	2.00	.00	.00	.20	.74	.05	.01	2.00
8389	3.00	1.00	1.00	.00	.21	.52	.26	3.00
9456	2.00	.00	1.00	.17	.76	.07	.01	2.00
etc.								

列 PRE_1 针对每个个案给出了具有某个因子（因子组合）最高期望概率的反应类别。对于 PATID 2364 和两个因子 KGEWICHT=0 和 EKINDER=1 的组合，第二个 BMI 变量类别是概率最大的类别，与观察到的数据 BMI=2（最左边）一致。

期望事件（在 PRE_1 一项下）不必与观察到的因变量类别相符。在 PATID 4795 的数据行中，尽管观察到的类别 BMI=4，但是将 BMI=3 预测为概率最大的类别。

变量 EST1_1 至 EST4_1 针对每个个案和所给的因子组合给出了期望概率。变量 EST1_1 至 EST4_1 针对每个个案始终得出概率为 1 或 100%。因此，变量 EST1_1 至 EST4_1 针对因子组合的等级始终是其各自预测的累积概率。对于 PATID 2364，组合 KGEWICHT=0 和 EKINDER=1（第二个 BMI 类别的概率，也就是 BMI=2）的变量 EST_2 具有最高的估计概率（0.76）。这个概率也可以表达为：一个具有特征组合 KGEWICHT=0 和 EKINDER=1 的人，以 76% 的概率归类到 BMI 类别 2。预测的类别就是概率最高的 BMI 类别，这个概率是以在相关个案的概率值为基础的。由于数据集含有 2×2 个独立类别，因此针对这些类别测定了所有四个从属类别都出现的各自概率，EST1_1 至 EST4_1 中的数据集最多含有 16 个不同的概率值。

预测优度检验

为了检查所建立模型的预测优度，最后使 BMI 和 PRE_1 之间相关起来。BMI 得到观察事件的类别，PRE_1 得到具有期望事件的类别。在理想情况下，两个变量之间的相关性应等于 1，从而可以推断出，以变量 PRE_1 为形式的模型正确无误地复制了变量 BMI。下面，利用肯德尔法使由模型预测的（PRE_1）类别与观察到的类别（BMI）相关起来。

相关性			BMI	预测的反应类别
肯德尔 τ -b 系数	BMI 类别	相关系数	1.000	0.660**
		显著性（双侧）		0.000
		N	659	659
	预测的反应类别	相关系数	0.660**	1.000
		显著性（双侧）	0.000	
		N	659	659

相关性 $r=0.66$ 。预测优度作为一种全局量度，可能由于其他原因而不能令人满意。

BMI 类别*预测的反应类别的交叉列表

			预测的反应类别		总数
			20-25	26-30	
BMI 类别	≤20	数量	58	0	58%
		占 BMI 类别的百分比	100.0%	0%	100%
		占总数的百分比	8.8%	0%	8.8%
	20-25	数量	238	78	316
		占 BMI 类别的百分比	75.3%	24.7%	100.0%
		占总数的百分比	36.1%	11.8%	48.0%
	26-30	数量	24	172	196
		占 BMI 类别的百分比	12.2%	87.8%	100.0%
		总数的百分比	3.6%	26.1%	29.7%
	>30	数量	0	89%	89%
		占 BMI 类别的百分比	0%	100%	100%
		占总数的百分比	0%	13.5%	13.5%
总数	数量	320	339	659	
	占 BMI 类别的百分比	48.6%	51.4%	100.0%	
	占总数的百分比	48.6%	51.4%	100.0%	

如果额外地将两个变量交叉列表，并且输出结果带有百分数，则两个现象非常明显：变量 PRED_1 只具有两个变量类别，而不是像变量 BMI 一样有四个变量类别。在变量 PRED_1 中缺失了单元格 ≤20 或 >30，准确地说，这些单元格的数值还可能被分配到了相邻类别。但是即使重建剩余的类别，也没有达到最佳的效果。例如，针对 BMI 类别“20-25”的表格单元格显示，只正确分配了大约 75% 的数值；相反，针对 BMI 类别“26-30”的表格单元格显示，只正确分配了大约 88% 的数值。

3.3.6 有序回归的前提条件

1. 有序回归假设了（至少）一个自变量（ X ）和因变量（ Y ）之间的因果模型。根据逻辑，从一开始就排除了伪回归，例如，身高对学习成绩的预测。在模型中只给出了重要的变量，不重要的变量也应删除。
2. 成对的测量值 x_i 和 y_i 必须属于同一个对象。换言之，所调查的特征是从一个样本的同一个元素中提取的。
3. 自变量和因变量理想地相关紧密。
4. 观察值个数大于所有类别水平的数量乘以 5。经验法则是，一次分析应含有的个案数是所有自变量和因变量类别等级的数量的 5 倍。因此，若这个分析的设计含有总共 8 个类别水平，则应至少有 40 个个案纳入分析，无论这个分析是带有一个或者多个预测变量（对于交互作用的分析需要更多变量）。如果多个预测变量等级的组合可能导致产生很多空白单元格，则既可以从模型中剔除不重要的预测变量，也可以将预测变量等级合并起来。为了保证数据完整性，应小心地将预测变量等级合并起来。模型幂次越大或者功效越小，就应有越多的个案。Hosmer & Lemeshow（2000，339-347）提出了一个公式，除了样本量之外还能给定模型的幂次和检验方向。例如，针对作为经验法则的某些平行分布，建议至少 $N = 10 * n$ 个协变量（对于目标事件的不对称变量类别，情况则更为困难）。结果可能是只针对感兴趣的目标事件抽取个案、普遍地升高个案数量 N 以及（或者）减少协变量数量。
5. 缺失数据（缺失值）。尤其是对于预测模型，缺失数据可能导致问题。预测模型的理想条件是不缺失任何数据。如果数据是完全随机缺失的，则具体的缺失程度决定了分析时还留有多少百分比的数据，这还可能会导致问题。如果通过合理的思考，发现缺失值以某种方式与目标变量相关，那么只要从模型中剔除了这些缺失值，模型的解释和建模就会产生问题。例如，（a）从建模角度通过一个指示缺失值的指标和（b）从重建角度分析缺失数值（Missing Value Analysis），就可以将缺失数据重新引入模型，但是只能在这个前提条件下：这些缺失数据的编码、重建和模型集成是合理且可追溯的（如 Schendera，2007）。如果缺失值集中在一个变量上，则或许也可以从分析中剔除这些缺失值。
6. 为了保证卡方检验（如 McFadden 量度，皮尔逊）的可靠性，应遵守卡方检验的要求。例如，最多 20% 的单元格具有低于 5 的期望频率。
7. 平行性假设：有序回归提出了平行性假设，即所有因变量类别的斜率是一样的。应利用卡方检验来检验这个假设。如果平行线检验呈显著性，则这个模型违反了斜率恒定的假设。原因主要可能是链接函数不适当、模型不适合或者因变量的定序有问题。
8. 因变量。在有序回归中，因变量是定序和离散尺度的。有序回归的一种变体形式是基于所推导出累积概率的广义线性模型。
9. 链接函数。根据因变量的分布，应确定一个适当的链接函数。如果各个类别是均匀和对称分布的，则应选择 Logit 函数。如果潜变量具有很多极端值，则逆柯西函数是适当的。如果更高或者更低的类别概率更大，则互补重对数函数或者负重对数函数是适当的。如果潜变量呈正态分布（也就是对称的），则适用 Probit 函数。

10. 散布问题（超散布性或者超聚集性，又称 Overdispersion 或 Underdispersion）。对于正确设定的模型，模型拟合优度（皮尔逊、误差）量度除以自由度数量应得出一个在 1 左右的数值。若这个数值远远超过 1，则表明存在超散布性；低于 1 则表明是很少出现的超聚集性。散布问题表明也许没有二元分布，并且散布问题是导致产生错误的主要标准误差。在分析实践中经常碰到的超散布性，是由于模型缺失重要的预测变量，或者必须转换这些预测变量，存在离群值或者选择了错误的链接函数而造成的。通过将协方差矩阵改变尺度可以对离散进行校正，但是只有在检验和排除了其他错误源之后才能实施。
11. 因变量的定序编码。应从形式和语义角度检验因变量的编码。编码在某些情况下可能对所取得的统计量施加影响，可能无法用准确的数值表达出定量关系。应至少检验不同编码的效应。应避免由于编码原因（不取决于预测变量）造成空白单元格。编码的有序距离也应能够从语义上描绘出可解释的和可按等级排列的强度（不必是语义上的等距离）。如果无法将编码按照语义上可明确解释的顺序排列，则应进行重新编码。
12. 自变量。自变量的编码不影响因变量类别的出现概率。但是，多个独立预测变量的变量类别的组合可能导致产生零频数，从而损害卡方检验的有效性。尤其是对于多个预测变量的组合，应避免编码的分级太细（对于定距协变量更应如此）。如果要考虑到不同预测变量（等级）之间的交互作用，则应通过 /LOCATION 将这些交互作用添加到模型中。例如，如果在另一个分类变量（但不是预测变量）的变量类别中的一个预测变量具有每次都不同的变异性，则可以将这个模型调节一个尺度分量（在 SPSS 中可以通过 /SCALE）。大致上可以用一个权重因子改写尺度分量。只有当证明原模型的成效不大，或者原模型对估计值的解释越来越复杂时，才应将尺度分量添加到模型中。
13. 在残差图中，预测误差在零周围随机地散布，前提条件是样本足够大。
14. 预测优度检验。一个模型在完成参数化之后，应检验其预测优度是否具有实际关联性。通过最佳的估计，模型应可以正确地重现大部分观察到的事件。对于每个有关的类别单元格，数值低于 90% 是不可接受的，根据应用领域不同，甚至可以提出更高的要求。如果在预测的反应变量中不缺少类别，则可以检验与 Cohen κ 系数符合的程度。

3.4 多项逻辑回归

多项逻辑回归（又称多分类逻辑回归）是二元逻辑回归（准确地说：其扩展形式）和有序回归之间的一种“过渡”形式，因为多项逻辑回归一方面可以调查具有两个或者多个变量类别的因变量，另一方面是无须考虑其极差信息。

多项逻辑回归的计算方法原则上与二元逻辑回归相同，使多项 Logit 模型与多元定类因变量相配合（在本章结尾部分，详细阐述了二元逻辑回归和多项逻辑回归之间的区别，在这里不再专门阐明）。与此相反，有序回归的计算方法是基于累积 Logit 模型，也就是基于完全不同的模型假设。

例如，当要根据预测变量将个案分类时，就使用多项逻辑回归。因此，多项逻辑回归提出

的问题与逻辑回归或者有序回归的类似，举例如下。

- 根据哪些实验室参数可以将患者归类到病症 A、B 或者 C？
- 根据资产负债表上的哪些参数，可以将一个银行客户归类为明显信贷资质良好、潜在信贷资质良好或者不具备信贷资质，等等。

在特定的前提条件下，多项逻辑回归的计算方法要优先于二元逻辑回归的方法。例如，当所有的预测变量都是定类的，或者连续的预测变量只假设了有限数量的数值时。

对于采用 0/1 编码的因变量和离散尺度的预测变量，LOGISTIC REGRESSION 和 NOMREG 得出同样的结果（例如，上文的 Wald 统计量），只是在 SPSS 过程命令特定的输出结果中有所区别。

对于采用 0/1 编码的因变量和定比预测变量，LOGISTIC REGRESSION 和 NOMREG 会得出不同的结果，因为 NOMREG 是基于单个个案，而 LOGISTIC REGRESSION 只处理成组数据。

因此，对于定比预测变量应优先使用 LOGISTIC REGRESSION。

在绝大多数是甚至完全是定类的预测变量时，应优先考虑 NOMREG。因此，首先介绍的例子只使用了分类预测变量；关于巢式病例对照研究特殊个案的例子，则只使用定量预测变量。

3.4.1 例子、界面选择和语法：主效应模型（二元因变量）

下面这个例子的目的一方面是简单地概述多项逻辑回归的计算和解释，另一方面也是表明，利用 SPSS 过程命令 NOMREG 逐步计算带有一个二元因变量的模型所得结果与利用 SPSS 过程命令 LOGISTIC REGRESSION 所得的计算结果是相同的，除了 SPSS 过程命令特定的输出结果之外。

从 SPSS 12 版开始，在 NOMREG 新增加的“stepwise function”也可以在一次多项逻辑回归时利用四个逐步法测定出，所给出的预测变量中哪些是最好的。

个案示例

对一群患者（分别具有不同大小的肿瘤）进行如下调查：三个参数“雌性激素受体”（分为变量类别“阳性”和“阴性”，变量 ERPSTAT）、“黄体酮受体”（分为变量类别“阳性”和“阴性”，变量 PRSTAT）和“有淋巴结”（分为变量类别“是”和“否”，变量 LYMPH）是否有可能对发现的“肿瘤大小”（分为变量类别“小（<2cm）”和“大（2~5cm）”，变量 TUMOKAT2）有影响。

这个例子有助于回答两个问题：两个患者群的参数 ERPSTAT、PRSTAT 或者 LYMPH 是否有区别，各个参数的影响分别有多大？

界面选择（示例）

在 SPSS 程序主界面选择以下菜单项：分析 → 回归 → 多项 logistic。



将变量 TUMOKAT2 拖入“因变量”窗口。通过“参考类别”将因变量的一个变量类别确定为参考类别（例如，“第一个”，必要时包括排列，如“升序”）。将变量 LYMPH、ERPSTAT 和 PRSTAT 拖入窗口“因子”。在“模型”（右上）一项下设置“向后步进”一种逐步法。

窗口“协变量”。这个窗口与这次分析无关。分析模型不含有连续变量。单击“继续”按钮。

子窗口“模型”。选定选项“使常数项进入模型”，确定了模型应含有一个截距（ b_0 ）。选择“步进/用户定义”作为变量选择的方法。选定变量 LYMPH、ERPSTAT 和 PRSTAT，然后将其拖入“步进进入的变量”一栏（“F”表示因子，“C”表示协变量）。在“方法”一项下给定变量选择方法“向后步进”。单击“继续”按钮。

子窗口“统计量”。在“模型”一项下调用“个案处理总结”、“仿 R^2 ”、“步骤总结”、“关于模型拟合的信息”、“单元格概率”、“分类矩阵”和“拟合优度”。在“参数”一项下调用“估计值”和“似然比检验”。针对“定义子总体”预先规定，给出的独立因子（第一个选项）定义了用于单元格概率和拟合优度检验的协变量结构。通过“置信区间(%)”规定置信区间。单击“继续”按钮。

子窗口“标准”。首先在“标准”一项下接受所有的预设置。例如，“最大迭代次数”（100）、“最大步骤对分”（5）、“对数似然性收敛性”（0）、“参数收敛性”（0.000001）、“delta”（0）和“奇异性容许误差”（0.00000001）。激活选项“从迭代中向前检查分离数据点情况”，接受预设置的数值（20）。单击“继续”按钮。

子窗口“选项”。在“离散尺度”一项下保留预设置“无”。在“步进选项”一项下预先给定标准（概率），根据这些标准自变量进入方程（预设置： $p=0.05$ ）或者从方程中删除（预设置： $p=0.10$ ）。单击“继续”按钮。

子窗口“保存”。在本例中，保存估计的反应概率、预测的类别及其预测概率和实际概率。单击“继续”按钮。

单击“确定”按钮开始计算。

语句:

NOMREG

```
tumokat2 (BASE=FIRST ORDER=ASCENDING) BY lymph erpstat prstat
/CRITERIA CIN (95) DELTA (0) MXITER (100) MXSTEP (5) CHKSEP (20)
      LCONVERGE (0) PCONVERGE (0.000001)
      SINGULAR (0.00000001)
/MODEL = | BSTEP = lymph erpstat prstat
/STEPWISE = PIN (.05) POUT (0.1) MINEFFECT (0) RULE (SINGLE)
/INTERCEPT = INCLUDE
/PRINT = CELLPROB CLASSTABLE FIT PARAMETER SUMMARY
      LRT CPS STEP MFI
/SAVE ESTPROB PREDCAT PCPROB ACPROB .
```

备注: NOMREG 命令调用计算多项逻辑回归的计算方法。接着给出一个多元（在这里是二元）尺度因变量（在这里是 TUMOKAT2）。因变量可以是定量的或者多元的（理想情况下较短的）字符串变量。

ORDER 和 BASE 首先共同定义了因变量数值的排序方式，然后定义了哪个数值表示参考类别（Base）。对于参考类别的模型参数不进行估计。ORDER= 确定了因变量数值的排序。利用 ASCENDING 或者 DESCENDING 将数值按升序或者降序排列。在采用 ASCENDING 时，最低值定义了第一个类别，最高值定义了最后一个类别；在采用 DECENDING 时，最高值定义了第一个类别，最低值定义了最后一个类别，通过 BASE=可以将一个不是因变量（预设置最后一个类别的类别设定为参考类别，然后通过 BASE 更精确地设定哪个数值应是参考类别。通过 DATA，由未分类数据的第一个数值构成第一个类别，由未分类数据的最后一个数值构成最后一个类别。借助于 BASE=DATA，得出的结果取决于数据集的分类，因此，不同的分类会得出不同的结果。BASE= 在通过 ORDER 定义的数值序列中确定了因变量的具体参考类别（“Basis”）。例如，在本例中，FIRST 将第一个（最低的）类别定义为参考类别。其他方法包括 LAST 命令或者给定一个数值代码，例如，以日期、货币或者字符串的形式。根据 BY 命令，给定模型的因子（在这里是 LYMPH、ERPSTAT 和 PRSTAT）。因子可以是数值或者（理想情况下比较短的）字符串变量。如果还要给定定距协变量，则这些协变量必须是数值，并且在语句中位于根据 WITH 命令预测的变量后面。

在 CRITERIA 一项下主要给出了估计算法的设置。例如，奇异性容许误差。利用 CIN (95, 预设置) 定义置信区间为 95%；在 CIN 一项下可以给定从 50 到 99.99 的数值。在 DELTA 一项下可以给定 0 到 1 之间的数值，然后将这个数值添加到空白单元格，从而确保了估计算法的稳定性。在本例中，空白单元格不应加入数值。DELTA 不应与 BIAS（不使用）混淆，在 BIAS 一项下可以类似地给定一个在 0 到 1 之间的数值，然后将这个数值添加给所有观察到的单元格频率。

MXITER (100) 和 MXSTEP (5) 给定了迭代最大次数和最大对分步骤（这些数值必须是正整数）。在达到设定的最大次数后，估计迭代停止。在 CHKSEP (20) 一项下可以给出一次迭代，在这次迭代时，计算方法要开始验证数据的完全分离或准完全分离（预设置为第 20 次迭代）。

通过 LCONVERGE 和 PCONVERGE，设置对数似然值函数和参数估计的收敛标准。在

LCONVERGE 一项下，给出达到对数似然值收敛的一个阈值，预设置为 0。以常规方式给出这个数值（与此相反，在使用 SPSS 过程命令 PLUM 时是采用科学计数法）。当最后两次迭代之间的对数似然值绝对或相对变化小于这个值时，迭代过程结束。在预设定为 0 时，不使用这个标准。在 PCONVERGE 一项下，给出达到参数收敛的一个阈值。以常规方式给定这个数值，预设置是 0.000001。如果参数估计值中的绝对或者相对变化小于这个值，则可以认为这个算法达到了正确的估计值。在 PCONVERGE (0) 时，不使用这个标准。在 SINGULAR 一项下，可以给定奇异性容许误差值，预设置是 0.00000001。

在 /MODEL 一项下给定模型效应，在一条垂直线（“|”）后面给定对其进行选择的方法。如果没有使用逐步法，则取消（“|”）和针对计算方法的选项。如果要计算一个全因子模型（子命令/FULLFACTORIAL），则取消根据/MODEL 的所有给定数据，见下文）。下面的阐述指的是根据/MODEL 给定的变量，然后是 BSTEP 一项下给定的可选择变量。

在 /MODEL 一项下，通过简单的预测变量给定模型主效应。例如，通过“(A*B)”或者“A BY B”形式的表达方式，给定主效应之间的交互作用。因子 A “嵌套”在因子 B 内部时产生的嵌套效应，例如，可以通过“A(B)”或者“A WITHIN B”形式进行设定，而且可以给定多重嵌套效应。为了使一个协变量进入模型（协变量指的是定比变量，其主要作用是帮助解释因变量的未检验方差），必须在/MODEL 一项下根据 WITH 命令给定这个协变量，协变量不能嵌套。如果没有根据 /MODEL 命令做出设定，或者退出了这个命令，则由根据“BY”（在 NOMREG 一项下）给定的预测变量创建一个标准模型。标准模型首先含有常数（如果已经进入模型），然后是按照给定顺序的协变量（如果含有的话），最后是按照给定顺序的预测变量。

如果确定了模型的效应，则在一条垂直线（“|”）后面给定对其进行选择的方法，可以在 BACKWARD、FORWARD、BSTEP 和 FSTEP 之间进行选择（详细情况参见下文）。如果没有使用逐步法，而是使用了一个直接的主因子模型，则取消（“|”）关于计算方法的说明。如果通过 /FULLFACTORIAL 调用一个具有预测变量所有可能交互作用的全因子模型，则完全取消 /MODEL 各行。MODEL 和 FULLFACTORIAL 命令，或者 MODEL 和 TEST 命令不能同时使用。

在 /MODEL 子命令中，不得多次对一个效应做出给定，例如，在给定其他命令（如 BSTEP）时。因此，无论是从理论上还是根据/MODEL 命令必须给定一个预测变量（因为其属于模型），也必须根据 BSTEP=给定这个预测变量（因为其属于需要步进选择的预测变量），但在命令行中只允许根据 BSTEP 给定一次这个预测变量（参见例子）。如果根据 MODEL 给定的预测变量少于根据 BY 给定的，则这些变量只用于定义子总体，不用于创建模型。

在本例中，通过 BSTEP（见下文）选择了“向后步进”法。在“=”符号后面列出了需要检验、进入模型或者从模型中剔除的变量（变量交互作用）等（在本例中是作为主效应的 ERPSTAT、PRSTAT 和 LYMPH）。从 SPSS 12 版开始，在 NOMREG 中，针对多项逻辑回归就可以使用变量选择下列逐步法。

BACKWARD（向后剔除）。第一步根据 BACKWARD 一次性将所有变量（或者嵌套效应、交互效应）进入模型，然后依次检验是否可以剔除。剔除似然比统计量显著性值最高（超过 PORT）的变量。重新检验剩下的模型。不断重复这个过程，直到不再有变量达到剔除标准为止。

FORWARD（向前进入）。根据 FORWARD，依次检验变量（或者嵌套效应、交互效应）是否可以进入模型。似然比统计量显著性值最低（小于 PIN）的变量进入模型，重新检验所产生的模型。不断重复这个过程，直到不再有变量达到进入标准为止。

BSTEP（向后步进）。第一步根据 BSTEP 一次性将所有变量（或者嵌套效应、交互效应）进入模型，然后依次检验是否可以剔除。剔除似然比统计量显著性值最高（超过 PORT）的变量。重新检验剩下的模型。不断重复这个过程，直到不再有变量达到剔除标准为止（在这一点上，BSTEP 法与 BACKWARD 法一致）。然后检验不在模型中的变量是否可以进入模型。似然比统计量显著性值最低（小于 PIN）的变量进入模型。重新检验所产生的模型。不断重复这个过程，直到所有变量满足进入或者剔除标准为止。

FSTEP（向前步进）。根据 FSTEP，依次检验变量（或者嵌套效应、交互效应）是否可以进入模型。似然比统计量显著性值最低（小于 PIN）的变量进入模型。重新检验所产生的模型（在这一点上，FSTEP 法与 FORWARD 法一致）。然后检验模型中现有的所有变量是否可以剔除。剔除似然比统计量显著性值最高（超过 PORT）的变量。重新检验剩下的模型是否可以剔除。只要不再有变量满足剔除标准，就重新检验变量是否可以进入模型。不断重复这个过程，直到所有变量满足进入模型的标准或者剔除标准为止。

如果给定了任意一个逐步法，则忽略子命令 TEST，也可能没有给定常数（截距）作为效应。给定的逐步法从根据/MODEL 给定因变量的结果开始，如果在这里没有给定效应，则初始模型就是常数模型（INTERCEPT=INCLUDE）或者零模型（INTERCEPT=EXCLUDE）。在显著性值都相同（例如，在 PIN 方面的数值相同）的情况下，首先让第一个给定的效应进入模型或者从模型中剔除。关于在调用逐步法时的细节和特殊性，请参见 SPSS 的技术文档。

根据/STEPWISE，可以对逐步法进行进一步设置。利用 PIN 和 POUT 确定，根据哪些基于似然比统计量的参数使变量进入模型或者从模型中剔除。针对 FORWARD、BSTEP 和 FSTEP，利用 PIN (.05) 预先设定了决定一个变量进入模型的数值。如果一个变量的概率小于进入值（如预设置为 0.05），则这个变量进入模型。给定的进入值（PIN）越大，变量就越容易进入模型。SPSS 中预设置的 0.05 被认为是相对保守的，为了使潜在的有关预测变量进入，最多可以接受 0.20。针对 BACKWARD、BSTEP 和 FSTEP，利用 POUT (0.1) 定义了决定从模型中剔除变量的数值。如果一个变量的概率大于剔除值（如预设置为 0.1），则从模型中剔除这个变量。给定的进入值（POUT）越大，变量就越容易留在模型中。进入值必须小于剔除值。

通过 MINEFFECT (0)（针对 BACKWARD 或者 BSTEP）或者 MAXEFFECT (0)（针对 FORWARD 或者 FSTEP），可以预先设定最终模型的效应的最大或者最小数量，常数（截距）不算作效应。对于 MINEFFECT，预设值为 0；对于 MAXEFFECT，这个值等于 NOMREG 中给定的所有效应的总数量。

通过 RULE，可以在括号内给定进入或者剔除模型元素的其他规则。所有规则（除了 NONE 之外）都用于确定等级。等级的条件是，只有当其元素（如 A 和 B）先前已经进入模型，更高的效应（如 A*B）才能进入模型。所有规则（SINGLE、SFATOR、CONTAINMENT 和 NONE）的共同点是，每次只能让一个效应进入模型或者从模型中剔除一个效应。这些规则的不同之处体现在对协变量的处理。根据模型不同，SINGLE、SFATOR 和 CONTAINMENT 尽管具有同样的效应，但有时会归入相同的等级，有时也会归入不同的等级（参见下面表格中的

示例)。SINGLE (预设置)的前提条件是,等级适用于模型中的所有效应;对协变量的处理和因子一样处理。SFACOR 的前提条件是,等级只适用于模型中的所有因子项,可以随时给定带有协变量的所有其他变量。CONTAINMENT (又称包含,进入)在协变量效应内部适用,前提条件是等级只适用于带有协变量的因子。对于 NONE,则既不需要给定等级,也不需要给定包含。详细情况可以参见 SPSS 语法文档和统计学专业文献。

效应 (举例)	方法		
	SINGLE	SFACOR	CONTAINMENT
A, B, A*B	1. A, B, 2. A*B	1. A, B, 2. A*B	1. A, B, 2. A*B
A, A**2, A**3	1. A, 2. A**2, 3. A**3	任意顺序	任意顺序
A, B, B (A)	1. A, B 2. B (A)	任意顺序	1. B, 2. B (A), A 的任意顺序

通过 /INTERCEPT INCLUDE, 可以使常数进入模型。通过将变量类别的数量减 1, 可测定出常数的数量。通过 EXCLUDE, 可以从模型中剔除常数项。

通过/PRINT 确定输出结果。CELLPROB 调用了一个观察频率和预测频率 (包括皮尔逊残差), 以及协方差模式和反应类别的观察百分比和预测百分比的表格。CLASSTABLE 调用了因变量观察值和预测值类别的一个表格, 呈交叉列表形式, 并且可以对模型的性能做出初步估计。FIT 调用关于模型拟合的信息, 并且利用皮尔逊卡方统计量或者 2*对数似然值表达了带有预测变量的模型是否能比纯粹的常数模型提供更好的信息。卡方显著性是一种符合期望的显著性。这些统计量是基于标准模型或者在 SUBPOP 一项下给定的变量。

子命令 PARAMETER 调用参数效应 (模型项) 的估计值。SUMMARY 调用 “拟 R^2 ” 表。从这个表中可以获取测定的拟 R^2 统计量 (Cox & Snell、Nagelkerke 或 McFadden) 作为模型归纳。 R^2 统计量近似地测量由模型解释的、因变量中的方差分量。 R^2 统计量越大 (只有在 Nagelkerke R^2 时最大值=1.0), 解释的方差分量越大 (关于这些统计量的详细情况参见有序回归一章)。

LRT 输出似然比检验。似然比检验检验了模型效应的系数是否显著区别于 0。这个表格含有模型的检验统计量和模型的偏效应。如果没有给定 LRT, 则只输出模型的检验统计量。CPS 展现了针对所给定变量的个案。STEP 针对逐步法的每个步骤归纳了进入模型或者从模型中剔除的预测变量 (步骤综述的输出结果只用于逐步法)。MFI 调用关于模型拟合的信息, 其形式是常数模型与最终模型的比较。

根据 /SCALE (不在本例语句中), 可以给定一个用于消除散布问题的校正系数。通过 DEVIANCE, 在使用离差函数 (似然比卡方) 的情况下估计尺度值。根据 Menard (2001) 的著作, 离差量度比皮尔逊量度的信息量更大。通过 PEARSON, 在试验皮尔逊卡方统计量的情况下估计尺度值。在 N 一项下, 也可以给定一个为正数的尺度值。改变尺度只涉及参数估计值 (如标准误差), 并不涉及对参数的估计 (如模型拟合, 拟 R^2 或者似然比检验)。根据经验, 只有在发现散布问题之后才能将 /SCALE 输入分析 (有些情况下从而使对估计值的解释更加复杂)。

通过 SAVE 命令, 可以逐个案地将估计统计量保存到活动数据集。通过 ESTPROB 可以

保存估计概率，即一个因子/协方差模式有多大概率被归入一个反应类别（即“估计反应概率”）。估计概率的数量和因变量类别的数量一样多。根据标准，保存了 25 个估计概率。通过 **PREDCAT** 可以保存具有因子/协变量模式最大期望概率的因变量类别（即“预测类别”）。通过 **PCPROB** 可以保存估计概率，即一个因子/协变量模式有多大概率可以被归入预测类别，这个概率等于估计类别概率的最大值（即“预测的类别概率”）。通过 **ACPROB** 可以保存估计概率，即一个因子/协变量模式实际上有多大概率可以被归入观察到的因变量类别（即“实际的类别概率”）。如果没有另做说明，则预先设定的关键词也是由 **SPSS** 创建的变量名的一部分。**ESTPROB** 创建 **est1_1**、**est1_2** 等；在活动数据集中显示为各自反应类别的“估计单元格概率”，如 1、2 等。**PREDCAT** 创建 **pre_1**（“预测的反应类别”），**PCPROB** 创建 **pcp_1**（“预测类别的估计分类概率”），**ACPROB** 创建 **acp_1**（“实际类别的估计分类概率”）。

其他选项（在本例中没有使用）如 **CORB** 和 **COVB**（渐近估计的成对相关系数或者成对协方差的矩阵，当有缺失值或者空白单元格时，其中一个变量是冗余的，或者两个变量都是冗余的）、**HISTORY**（每一次迭代的所有函数值和参数估计值的表格形式列表）和 **KERNEL**（只调用对数似然值函数的核心，而不是完整的对数似然值函数，也就是说没有多项常数）。

通过 **SPSS** 语句，在上述的程序示例中还可以设置更多的输出结果。为了计算一个多项逻辑回归，可以根据自己的要求设置很多选项（如 **FULLFACTORIAL**、**OUTFILE**、**SUBPOP** 或 **TEST**）来进一步了解所介绍的 **NOMREG** 示例。详细情况可以参见 **SPSS** 语法文档和统计学专业文献。

3.4.2 输出结果和解释

名义回归

在这个标题后面，是一个多项逻辑回归的输出结果。

个案处理个案		数量	边界百分比
肿瘤大小	小 (<2cm)	369	59.6%
	大 (2-5cm)	250	40.4%
淋巴结	否	160	25.8%
	是	459	74.2%
雌性激素受体	阴性	241	38.9%
	阳性	378	61.1%
黄体酮受体	阴性	274	44.3%
	阳性	345	55.7%
有效个案		619	100.0%
缺失个案		243	
总个案		862	
子总体		8	

针对每次分析，表“个案处理总结”给出了有效个案的总数（ $N=619$ ）和缺失个案的（ $N=243$ ）总数。与二元逻辑回归相反，有效个案的表现是每次根据预测变量（**LYMPH**，

ERPSTAT, PRSTAT) 和因变量 (TUMOKAT2) 的变量类别来展示的。通过所有预测变量的所有相乘的变量类别而测定出“子总体”。由于所有三个预测变量都有两个可能取值, 因此结果是 8。例如, 可以从“观察频率和预测频率”(见下文) 中查出, “子总体”起到什么样的作用。

模型拟合信息

模型	操作	效应	2*对数似然值	卡方	自由度	显著性
步骤 0 0	已输入	<所有> ^a	38.705			
步骤 1 1	已删除	prstat	39.113	.408	1	.523

逐步法: 向后步进

a. 这个模型含有在子命令 MODEL 中明确或者隐含给定的所有效应。

在“步进纵览”表下面给定了所设置的逐步法, 例如, 在这里是“向后步进”。显示的内容取决于所选择的方法和预设置。在步骤 0 一行, 给定了初始模型 (输入了所有效应) 的 2*对数似然值。在步骤 1 一行, 首先显示过程 (“操作”), 然后显示涉及的变量 (在这里是 PRSTAT) 及其参数: 2*对数似然值、卡方值、自由度和显著性。在本例中, 在步骤 1 时, 利用设置的方法“向后步进”从模型中剔除了显著性水平最高 (超过 PORT) 的变量 (在这里是 PRSTAT)。重新检验剩余的模型。本例在步骤 1 之后停止。所有剩余的变量 (在这里: LYMPH, ERPSTAT) 符合进入标准或者剔除标准。

模型拟合信息

模型	2*对数似然值	卡方	自由度	显著性
只有常数项	69.655			
最终	39.113	30.542	2	.000

表“模型拟合信息”表明了模型是否有能力做出精确的预测。“卡方”值 (30.542) 是基于常数模型和最终模型的 2*对数似然值之间的差值, 并检验带有预测变量的模型 (包括预测变量 ERPSTAT 和 LYMPH 在内的总模型, “最终”, 39.113) 是否能比没有预测变量的模型, 也就是常数模型 (“只有常数项”, 69.655) 提供更好的信息。针对常数模型 (“只有常数项”) 和总模型 (“最终”) 输出 2*对数似然值。如果在模型中出现很多空白单元格 (零频数), 则对数似然值数据本身可能就是值得怀疑的; 但是, 对数似然值之间的差值通常总是解释为 (接近于) 卡方分布的 (McCullagh & Nelder, 1989)。卡方的显著性 ($p=0.001$) 表明, 带有预测变量的模型能够比纯粹的常数模型提供更好的信息。卡方显著性是一种符合期望的显著性。下面这个关于模型拟合优度的表格介绍了总模型在多大程度上得到了改善。

拟合优度

	卡方	自由度	显著性
皮尔逊	5.332	5	.377
离差	4.590	5	.468

表“拟合优度”表明，实际观察的单元格频率是否和在多大程度上与利用模型计算出的期望概率有显著区别。卡方检验（皮尔逊，离差）的 p 值为 $p=0.377$ 或 0.468 ，因此不会得出显著的结论，这就表明计算出的模型具有很高的拟合优度。在做这些检验时，不希望达到显著性，因为这表明模型与数据有统计学上的重大偏差。根据 Menard (2001) 的著作，对于逻辑回归而言，离差量度比皮尔逊量度具有更多的信息量。对离差量度不利的是，离差量度不仅有对自变量的辨别能力，而且还对基本比率（如不对称地占满因变量类别的单元格）做出反应。当占满的单元格很少或者甚至空白的单元格很多时，使用卡方检验是有问题的。这些拟合优度量度表明存在超散布性或者超聚集性。卡方值离差（4.590）除以自由度（5）得出的是上一个接近于 1 的数值（0.918）。没有超散布性或者超聚集性的迹象。

拟 R^2	
Cox & Snell	0.048
Nagelkerke	0.065
McFadden	0.037

从“拟 R^2 ”统计量表可以查出，通过模型解释的方差占多大比例。目标值约为 1.0 或 100%。Nagelkerke R^2 表明模型解释了大约 6.5% 的方差。而且 McFadden R^2 也保持低于 0.20，这表明模型拟合良好。对于多项逻辑回归，可以不用精确地，而是只以近似值的形式计算出 R^2 统计量。

所实施多项逻辑回归的主要结果可查看“参数估计值”表。

似然比检验				
效应	2*对数似然值	卡方	自由度	显著性
常数项	39.113 ^a	.000	0	.
lymph	52.667	13.554	1	.000
erpstat	54.261	15.148	1	.000

卡方统计量展现了最终模型和缩减模型的 2*对数似然值之间的差值。通过从最终模型中删除一个效应，计算出缩减模型。在这里是以零假设为基础，根据这个零假设，这个效应的所有参数为 0。

a. 这个缩减模型等同于最终模型，因为删除一个效应并没有增加自由度的数量。

似然比检验是检验各个预测变量是否对模型起作用。从最终模型中删除经过检验的预测变量，从而计算出缩减模型的 2*对数似然值。当有一个显著的预测变量时，卡方值等于缩减模型和最终模型的 2*对数似然值之间的差值（参见 LYMPH）。似然比检验被视为超过 Wald 检验（Hosmer & Lemeshow, 2000; Menard, 2001, SPSS, 2003）。似然比检验更加精确，此外，当回归系数较大时，Wald 统计量会导致 II 类错误。只要预测变量的显著性低于（例如）0.05 的 Alpha 值，这个预测变量就对模型起作用。因此，无论是 LYMPH 还是 ERPSTAT 都对模型起作用（每次 $p=0.000$ ）。

参数估计值

肿瘤大小 ^a		B	标准误差	Wald 统计量	自由度	显著性	Exp(B)	EXP(B)的 95%置信区间	
								下限	上限
大 (2~5cm)	常数项	.036	.113	.103	1	.748			
	[lymph=0]	-.727	.203	12.877	1	.000	.483	.325	.719
	[lymph=1]	0 ^b	.	.	0
	[erpstat=0]	-.676	.176	14.752	1	.000	.509	.360	.718
	[erpstat=1]	0 ^b	.	.	0

a. 参考类别为：小（< 2cm）。

b. 由于这个参数是冗余的，因此设为零。

表“参数估计值”针对所计算模型的预测变量汇总了 **B**（回归系数，非标准化）、标准误差、Wald 统计量、自由度、Wald 显著性、Exp(B)（胜率）和胜率的置信区间。Wald 统计量是基于各个参数估计值与其标准误差的商的平方。如果变量具有两个变量类别，则只针对参考类别输出 Wald 统计量。对于参数的解释，重要的是显著性和正负号。只有当一个参数的显著性低于（例如）0.05 的 Alpha 值时，这个参数才区别于 0，并且对模型有作用。所调用的 Exp(B) 置信区间不应包括 1。在模型中，只留下了分类变量。对其参数的解释与定量变量的解释只有细微区别（参见 Menard，2001）。为了能够解释这个表格，变量的编码必须是已知的。在括号内给出了所指的类别编码，因此在 NOMREG 时，指的就是用户定义的编码（与 LOGISTIC REGRESSION 相反，见下文）。最后一个类别是冗余的，是根据其他分组得出的（如果在回归方程中没有包含常数，则对于最后一个类别来说常数不是冗余的）。具体而言，这就表示无须明确的编码，而只是利用常数，即没有回归系数，就可以计算出最后一个类别的概率。

与各自的参考类别相比，具有正系数（参见“**B**”）的参数提高了因变量各自类别的概率，具有负系数（参见“**B**”）的参数降低了因变量相应变量类别的概率。在本例中，LYMPH 和 ERPSTAT 分别得出统计学上重要的显著性（0.000）。这两个变量都是统计学上的重要预测变量（但是从内容上来看是否也重要，则完全是另外一个问题）。可以十分简便地通过胜率 Exp(B) 看出分类预测变量的相对意义（B 是非标准化的，可能有很大的误导作用）。Exp(B) 的数值超过 1 越多，各自变量的影响也就越大。例如，LYMPH 的数值（B=-0.727，（Exp(B)=0.483）应解释为：在 LYMPH 值为 0（“否”）时，因变量的类别“大（2~5 cm）”的概率与参考类别“小（< 2 cm）”（参见表格脚注）的概率相比，小了大约 50% 或者 2 倍。反之，如果是 LMYPH=1，则这个结果应解释为：在淋巴结一项为“是”（LYMPH=1）时，出现相依类别“大（2~5 cm）”的概率与参考类别相比大约是其 2 倍。对于 ERPSTAT 得出类似结果。

分类

观察值	预测值		
	小（<2cm）	大（2~5cm）	正确率百分比
小（<2cm）	228	141	61.8%
大（2~5cm）	101	149	59.6%
正确率百分比	53.2%	46.8%	60.9%

表“分类”是因变量观察值和预测值的一个交叉列表。在从左上到右下对角线上的数值表示正确的预测。各个错误分类位于对角线的旁边。有 60.9% 是正确预测的值。这个模型的性能不是很强。

观察频率和预测频率

黄体酮受体	雌性激素受体	淋巴结	肿瘤大小	频率			百分比	
				观察	预测	皮尔逊残差	观察	预测
阴性	阴性	否	小 (<2cm)	48	46.217	0.582	82.8%	79.7%
			大 (2~5cm)	10	11.783	-0.582	17.2%	20.3%
		是	小 (<2cm)	87	85.098	0.351	66.9%	65.5%
			大 (2~5cm)	43	44.902	-0.351	33.1%	34.5%
	阳性	否	小 (<2cm)	8	8.661	-0.389	61.5%	66.6%
			大 (2~5cm)	5	4.339	0.389	38.5%	33.4%
		是	小 (<2cm)	36	35.839	0.038	49.3%	49.1%
			大 (2~5cm)	37	37.161	-0.038	50.7%	50.9%
阳性	阴性	否	小 (<2cm)	8	11.156	-2.096	57.1%	79.7%
			大 (2~5cm)	6	2.844	2.096	42.9%	20.3%
		是	小 (<2cm)	25	25.529	-0.178	64.1%	65.5%
			大 (2~5cm)	14	13.471	0.178	35.9%	34.5%
	阳性	否	小 (<2cm)	52	49.966	0.498	69.3%	66.6%
			大 (2~5cm)	23	25.034	-0.498	30.7%	33.4%
		是	小 (<2cm)	105	106.534	-0.208	48.4%	49.1%
			大 (2~5cm)	112	110.466	0.208	51.6%	50.9%

表“观察频率和预测频率”中含有有效个案 ($N=619$) (包括皮尔逊残差在内) 及预测变量 ERPSTAT、LYMPH 和 PRSTAT 的协变量模式的观察百分比和预测百分比, 以及因变量 TUMOKAT2 的类别的观察百分比和预测百分比。在右下角的百分比下面, 观察值大致等于预测值, 子总体“黄体酮: 阳性”+“雌性激素: 阴性”+“淋巴结: 否”除外 (对此参见皮尔逊残差)。必要时, 可以通过让一个预测变量进入从而改进这个子总体相对较差的预测。在本例中, 这个矩阵的结构是基于主效应, 但是可以通过 SUBPOP 命令单独设置。如果子总体吸纳了一个或者多个定量协变量, 则当这些协变量导致空白单元格时, 模型拟合的量度就不适合使用了。如果空白单元格的比例上升大约 5% (经验法则), 或者集中到内容最重要的单元格上, 则表明出现了统计学上的问题, 就如内容上的问题一样。

3.4.3 补充说明: 逐步计算带有一个二元因变量的模型: NOMREG REGRESSION 和 LOGISTIC REGRESSION 输出结果的比较

如果一个模型带有一个二元因变量和同一组分类预测变量, 或者变量选择时采用同样方法的模型, 则对其计算会得出同样的结果。前面介绍的例子也用 LOGISTIC REGRESSION 做了计算。为了阐明起见, 下面列出主要的输出结果以做比较。

表“模型总结”基本上与 NOMREG 的“拟 R^2 ”表格一致。

模型总结

步骤	2*对数似然值	Cox & Snell R ²	Nagelkerke R ²
1	804.146 ^a	0.049	0.066
2	804.554 ^a	0.048	0.065

a. 估计过程在第 4 次迭代时结束，因为参数估计值的变化幅度小于 0.001。

拟 R²统计量（Cox & Snell, Nagelkerke）完全一致。但是代替 McFadden R², LOGISTIC REGRESSION 输出了一个步进测定的 2*对数似然值。

表“方程中的变量”大致上等同于 NOMREG 的“参数估计值”。测定的参数（如回归系数 B、Wald 等）完全一致。在 NOMREG 计算时，利用 BASE=FIRST 使参数估计值针对的是同一个参考类别。在括号中的这两个计算方法的结果表格表明，估计值针对的是各个变量的哪种编码。

方程中的变量

	回归系数 B	标准误差	Wald 统计量	df	显著性	Exp(B)	EXP(B)的 95%置信区间	
							下限值	上限值
步骤 1 ^a lymph (1)	-.734	.203	13.073	1	.000	.480	.322	.714
erpstat (1)	-.606	.207	8.583	1	.003	.546	.364	.818
prstat (1)	-.128	.200	.408	1	.523	.880	.594	1.303
Konstante	.067	.123	.299	1	.585	1.069		
步骤 2 ^a lymph (1)	-.727	.203	12.877	1	.000	.483	.325	.719
erpstat (1)	-.676	.176	14.752	1	.000	.509	.360	.718
Konstante	.036	.113	.103	1	.748	1.037		

a. 在步骤 1 输入的变量: lymph、erpstat、prstat。

与 LOGISTIC REGRESSION 相反，NOMREG 不输出表格，从这个表格可以获取分类变量的 SPSS 内部编码。在括号内，给定了每次使用的类别编码。在运行 LOGISTIC REGRESSION 时，采用了内部由 SPSS 分配的编码（“0”）。在运行 NOMREG 时，采用了外部由用户定义的编码（“1”）。由于在运行 LOGISTIC REGRESSION 时，内部由 SPSS 分配的编码（“0”）与在运行 NOMREG 时，外部由用户定义的编码（“1”）输出的参数估计值完全一致。一个细小的区别可能是在运行 LOGISTIC REGRESSION 时，输出结果的步进方式。

3.4.4 特殊情况：带有定量预测变量的巢式病例对照研究（1:1）——示例、语法、输出结果和解释

实施下文介绍的巢式病例对照研究是基于病例个案和对照个案之间的差异（参见 Hosmer & Lemeshow 的著作，2000，尤其是关于其他变体的第 6.3 节和第 7 章）。

ID	FALL	ALTER	SEX	SYS_T
1	1	34	0	4
2	1	53	1	3

3	0	34	0	6
4	0	53	1	5
...				

通常，一行中的一个数据组含有一个病例的数据（例如一个人的）。例如，第二行含有标为 ID= 2 的一个人，表示“病例”的编码“1”等。

分别用相同的变量调查病例组和对照组的人，例如，变量 ALTER、SEX 和 SYS_T。令人真正感兴趣的变量是 SYS_T；但是，同时人们还需先排除变量 ALTER 和 SEX 的效应。现在为了进行巢式病例对照研究而准备数据组，从而使由病例个案和对照个案构成很多对，并保存在单独一行中。构成一对的方式是，这一对的两个元素在潜在高杠杆变量（干扰变量）中是相同（配对）的，例如，年龄、性别等。因此，只有当来自病例个案和对照个案的元素（如 CASSYS_T 和 CONSYS_T）在成问题的配对变量（如年龄（ALTER）和性别（SEX））中具有相等数值时，才能将这些元素归在一行中。在任何其他情况下，都不能实现配对。如果人们对分析变量可能具有的区别感兴趣，则不将这些变量配对，而是分别针将病例个案或者对照个案保存为独立变量。表示病例个案的、SYS_T 中的数值被保存为 CASSYS_T，表示对照个案的数值被保存为 CONSYS_T。

巢式病例对照研究的数据集在每一行含有成对测量值的数据，在本例中也就是两个人。例如，第二行含有由标为 ID 2（病例个案）和 4（对照个案）的人组成的一对。

CASE_CON	CASE	CONTROL	ALTER	SEX	CASSYS_T	CONSYS_T
1_3	1	0	34	0	4	6
2_4	1	0	53	1	3	5
...						

在最终产生的数据集中，病例个案和对照个案在潜在高杠杆变量（干扰变量）中是相同（配对）的，例如，在 ALTER 和 SEX 中。这表示，这些变量的效应通过这次配对相同地分配给了这两个组（平行化），因此可以作为（统计上的）效应（干扰效应）予以排除。得到由配对数据组成的文件的最简单途径，是将病例个案和对照个案的数据保存在两个单独的数据集中，重命名其中固有的分析变量（因为之后的合并），然后以相同的方式将这两个文件排序，通过适当的键变量重新合并，最后将在合并过程中测定的成对数据纳入分析。这个方法主要是适用于测量值变异很小的数据，例如，定类数据。对于定距数据更适应其变体形式，尤其是当这些变体形式基于迭代优化的搜索和替代原理时，可以设置定距数据的“命中”精度。

如果对真正的分析变量之间的区别感兴趣，则在下一步通过简单地利用减法予以测定（而且还测定 0/1 编码的变量，并且包括因变量的编码）。

```
MATCOL = CASE - CONTROL.
DIFF = CASSYS_T - CONSYS_T.
```

对于变量 MATCOL，只有在为了实现控制时才利用减法。对于检验原本提出的问题，这个变量没有意义。如果所有步骤都执行正确，则 MATCOL 得出一个常数，这个常数假设了哪个具体数值是不重要的。下面给出了关于计算和解释的建议。变量 DIFF1...n 是在变量（干扰变量）中配对（相等）的皮尔逊成对数据的成对差值。因此，如果在结果中显示出了差值，则这些差值仅仅是由于归属于病例组或者对照组而造成的。只能用 SPSS 过程命令 NOMREG 分

析这个巢式病例对照研究，但是不能用 LOGISTIC REGRESSION，因为在这里因变量必须精确地具有两个变量类别。

* 利用 SPSS 的一个语句实例，前提条件是有一个配对的数据集*。

```
compute matcol = case - control .
exe.
compute diff1 = cassys_t - consys_t .
exe.
compute diff2 = casdia_t - condia_t .
exe.
compute diff3 = cassys_v - consys_v .
exe.
```

利用减法主要是对参数 (B)、Exp(B) 和置信区间的正负号有影响。为了之后在解释测定结果时能够更好地进行控制，建议通过将相应的病例值减去一个对照值，从而计算出对照个案与病例个案的偏差。因此，CONTROL 变成参考类别。另一种控制方法是将平均差值（如 DIFF2 的平均值）与各自的 Exp(B)（如 DIFF2 的胜率）进行比较。平均值和胜率的正负号必须一致。再将其推导过程，也就是具体的减法公式纳入考虑范围，就可以明确地确定效应的方向。对于因变量，对照个案 (CASE) 应编码为 0，病例个案 (CONTROL) 应编码为 1。然而对于 MATCOL，一个具体的差值应只能得出数值 1。这个数值只用于对照个案。常数的具体类别对结果没有影响。

```
NOMREG
matcol (BASE=LAST ORDER=ASCENDING) WITH diff1 diff2 diff3
/CRITERIA CIN (95) DELTA (0) MXITER (100) MXSTEP (5) CHKSEP (20)
LCONVERGE (0) PCONVERGE (0.000001) SINGULAR (0.00000001)
/MODEL
/STEPWISE = PIN (.05) POUT (0.1) MINEFFECT (0) RULE (SINGLE)
/INTERCEPT =EXCLUDE
/PRINT = CLASSTABLE FIT PARAMETER SUMMARY LRT CPS STEP MFI .
```

本语句在很大程度上与前面的示例一样。因此，在这里只指出了与分析有关的特殊性：尽管因变量 MATCOL 只具有一个变量类别，但是用 NOMREG 可以进行巢式病例对照分析。预测变量是定量变量，因此，在 /MODEL 一项下根据 WITH 命令给定这些预测变量。由于没有明确地预先设定逐步法，因此在本例中计算了一个直接的主效应模型。在 /INTERCEPT 一项下必须给定 EXCLUDE，否则无法执行这个分析。在 /PRINT 一项下应没有给定 CELLPROB，因为定量变量的观察频率和预测频率的列表是需要大量的、茫无头绪的并且在大多数情况下毫无效果的计算。

输出结果和解释

同样，输出结果和解释在很大程度上与前面的例子一样。下面只指出这次分析特有的特殊性：

警告

因变量只有一个有效数值。拟合一个有条件的逻辑回归模型。

测定的回归方程不含有常数。

经过处理的个案		
	数量	边际百分比
matcol -1	160	100.0%
有效个案	160	100.0%
缺失个案	0	
总体	160	
子总体	98	

表标题“经过处理的个案”是有误导性的，不是含有 160 个个案，而是含有 160 对测量值，也就是 320 个个案的数据。这么多数量的子总体是由于定量分析变量（协变量）的不同变量类别而造成的。

模型拟合优度			
	卡方	自由度	显著性
皮尔逊	219.041	95	.000
离差	118.956	95	.049

表“模型拟合优度”中的量度是不允许使用的。这些量度是基于因变量 MATCOL 的观察频率和预测频率，但是，由于 MATCOL 实质上是一个只有一个变量类别的常数，因此作为其基础的卡方频率表格，并不能提供足够多的基本数据（对此也可参见下文的表格“分类”）。此外，定量预测变量还会造成很多空白单元格。

模型拟合信息				
模型	模型拟合标准	似然比检验		
	2*对数似然值	卡方	自由度	显著性
零	221.807			
最终	118.956	102.851	3	.000

表“模型拟合信息”中的量度是允许使用的，最终模型与零模型具有统计学上的重要区别（ $p=0.000$ ）。

拟 R^2	
Cox & Snell	0.474
Nagelkerke	0.632
McFadden	0.464

达到 0.464 的 McFadden R^2 拟合得很好。模型解释了大约 63% 的方差（根据 Nagelkerke）。

参数估计值

matcol	B	标准误差	Wald 统计量	自由度	显著性	Exp(B)	EXP(B)的 95%置信区间		
							下限	上限	
1	diff1	-.105	.049	4.540	1	.033	.901	.818	.992
	diff2	.366	.081	20.207	1	.000	1.442	1.229	1.692
	diff3	.131	.038	11.957	1	.001	1.140	1.058	1.228

所有三个差值都是统计学上的重要预测变量。胜率（Exp(B)）的置信区间在任何情况下都不包含数值 1，所有预测变量都以 95%的概率向各组之间的关系施加影响。这个影响朝向哪个分析，主要是取决于做减法的方向（参见上文）。应始终用测定正负号的具体公式来校正测定的正负号，以便可以恰当地解释所得出的参数和排除特殊性（例如，双重否定）。在本例中，DIFF1 的负号表明，在变量 SYS_T 中的对照个案比病例个案具有更高的数值。DIFF2 和 DIFF3 的正号表明，病例个案平均比对照个案具有更高的数值，因此在预测变量数值增大时，“病例”类别的概率也随之增大。SPSS 不输出标准化系数，因此既无法绝对可靠地，也无法相对可靠地估计 DIFF1、DIFF2 或 DIFF3 的系数 B。Exp(B)在此可以起到指明方向的作用。例如，DIFF2 的胜率（Exp(B)=1.442）应解释为：在上升了一个单位的 DIFF2 时，因变量的比值也随之增加。因此，如果因变量中的比值先前是 1：1，则预测变量“DIFF2”升高一个单位会导致比例变成 1：1.442。如果 DIFF2 升高一个单位，则“病例”组的概率比“对照”组的概率增加了 44.2%，或者说达到 1.442 倍。

分类

观察	预测	
	1	正确率百分比
1	160	100.0%
总百分比	100.0%	100.0%

表“分类”的效果不佳。

3.4.5 补充说明：LOGISTIC REGRESSION 对比 NOMREG（区别）

SPSS 对于两种“同一个来源”的计算方法分别提供了一个单独的语句，根据经验，这不仅会造成误解，而且还会使看起来相等的模型拟合优度量度基于不同的计算方法（逐个案，逐组）。如果对于某个计算方法要使用另一种计算方法或者另一个语句的量度，这就增大了困难。

根据个案数据，SPSS 过程命令 LOGISTIC REGRESSION 针对二元逻辑回归测定了所有预测、残差、影响统计量和拟合优度检验。SPSS 过程命令 NOMREG 针对多项逻辑回归将个案汇总成组，从而在汇总数据的基础上构成带有相同的预测变量协变量模式的子总体。协变量模式可以想象成自变量的变量类别组合，其最简单的形式是 2×2 表格，自变量的数值分布在这个表格的单元格中。根据这个子总体做出了预测、残差和拟合优度检验。因此，两个回归方法的区别在于拟合优度检验的基本数据。模型拟合的分析不是根据因变量的变量类别数量区分的，而是根据个案（逐个案）或者协变量模式（成组）区别的。

如果个案数量超过协变量模式的数量,则应始终成组地测定模型参数,例如,通过皮尔逊或者 NOMREG 中的偏差。如果协方差模式的数量大致等于个案的数量,则应始终逐个案地测定模型参数(例如,通过 Hosmer-Lemeshow 检验,只有在 SPSS 过程命令 LOGISTIC REGRESSION 中可以使用,因此,多项模型必须分解为二项模型)。

关于 SPSS 过程命令 NOMREG 和 LOGISTIC REGRESSION 在计算方法和性能方面的其他共同点和区别,请参见第 3.4.4 节的一览表,以及 SPSS 技术文档和统计学专业文献。

3.4.6 多项逻辑回归的前提条件

1. 多项逻辑回归假设了(至少)一个自变量(X)和因变量(Y)之间的因果模型。根据逻辑,从一开始就排除了伪回归,例如,身高或者性别对头发颜色的影响。在模型中只给出了重要的变量,不重要的变量也应删除。
2. 成对的测量值 x_i 和 y_i 必须属于同一个对象。换言之,所调查的特征必须是从一个样本的同一个元素中提取的。
3. 自变量和因变量理想地相关紧密。
4. 因变量。在二元逻辑回归中,因变量是二分的;在多项逻辑回归中,因变量超过两个类别。对于多项逻辑回归而言,重要的是检验令人感兴趣的目标事件是否(频率足够地)出现,以及目标事件的频率究竟是符合总体,还是呈现出不平衡。如果目标事件很少,则通常假阴性的成本要超过假阳性的成本,并且分界值远远低于 0.5。
5. 缺失数据(缺失值)。尤其是对于预测模型,缺失数据可能导致问题。预测模型的理想条件是不缺失任何数据。如果数据是完全随机缺失的,则具体的缺失程度决定了分析时还留有多少百分比的数据,这也可能会导致出现问题。如果通过合理的思考,发现缺失值以某种方式与目标变量相关,那么只要从模型中剔除了这些缺失值,模型的解释和建模就会产生问题。例如,(a)从建模角度通过一个指示缺失值的指标和(b)从重建角度分析缺失数值(Missing Value Analysis),可以将缺失数据重新引入模型。但是只能在这个前提条件下:这些缺失数据的编码、重建和模型集成是合理和可追溯的。如果缺失值集中在一个变量上,则或许也可以从分析中剔除这些缺失值。
6. 多项逻辑回归假设了因变量和自变量之间的非线性函数,以及胜率对数的线性,即连续预测变量和因变量胜率对数之间的线性关联,这些都可以想象成一个线性的散点图。但是,目前在 SPSS 中还不能方便地输出 Logit 图。一种检验胜率对数线性假设的简便方法是,用每个连续预测变量及其固有算法之间的交互作用项补充初始的回归模型(称之为 Box-Tidwell 转换)。如果这些交互作用项其中的一个呈现出显著性,则这个模型就违反了胜率对数的线性假设,以及单调关联的假设(关于忽视非单调关联的风险请参见 Böhning 著作,1998,第 6 章)。这个方法的弱点是,无法识别轻微偏离线性的现象,并且在达到显著性时无法反映出非线性的形状。Menard (2001)介绍了其他的检验方法。非线性不能与非叠加性相混淆。
7. 模型的叠加。如果因变量相对于自变量的数值发生了达到自变量一个单位的变化,则出现非线性。与此相比,如果因变量相对于其中一个其他自变量的数值发生了达到自变量一个单位的变化,则呈现非叠加性。例如,可以通过检验是否存在可信的或者理

论上可能有的所有交互作用，检验模型的叠加性。后一种方法只适用于相对简单的模型。

8. 散布问题（超散布性或超聚集性，又称 *Overdispersion* 和 *Underdispersion*），对于正确设定的模型，模型拟合优度（皮尔逊，误差）量度除以自由度的数量应得出一个在 1 左右的数值。若这个数值远远超过 1，则表明存在超散布性；低于 1 则表明是很少出现的超聚集性。散布问题主要是在成组分析数据时出现，经常会导致标准误差错误。在分析实践中经常碰到超散布性的原因是，模型缺失重要的预测变量或者必须转换这些预测变量、存在离群值或者分布情况与假设的不一样。通过将协方差矩阵改变尺度可以对离散进行校正，但是只有在检验和排除了其他错误源之后才能实施。
9. 参考类别。参考类别对所测定结果的精度和方向具有决定性影响。例如，胜率编码为 1 时可能数值达到 3，但是编码为 0 时可能达到 0.33。例如，对于二元因变量（B）的系数，正负号发生改变。SPSS 过程命令 *NOMREG* 标准化地选择因变量的最后一个或者最高的变量类别作为参考类别（*BASE=LAST*，但是从 SPSS 12 版本开始，可以单独给定参考类别）。SPSS 过程命令 *LOGISTIC REGRESSION* 始终选择因变量的第一个或者最低的变量类别作为参考类别（在出现事件时编码为 1）。其他作者、分析师或者软件在有些情况下使用了其他的参考类别。请检查（自动）选择的参考类别是否符合评估目的。在临床或者流行病学研究中对于二元事件适用的规则是，始终将病例个案（暴露，事件）编码为“1”，始终将对照个案（不暴露，事件不出现）编码为“0”。如果在多项分析时，对概率较高的类别配较高的编码，这样就方便了对结果的解释。如果 *BASE=* 和/或 *ORDER=* 还不够，则可以通过 *RECODE* 将因变量的变量类别重新编码。
10. 自变量。预测变量应不相关（消除多重共线性）。预测变量之间的任何相关（如 > 0.80 ）都是多重共线性的迹象。通过容差检验，或者参数估计值明显很高的标准误差（非标准化： >2 ，标准化： >1 ）显示出多重共线性。通过针对同一个模型计算一次逻辑回归，就可以测定出容差量度（这个处理是允许的，因为只测定了预测变量之间关联的容差量度，因变量与此无关）。是否可以和在多大程度上可以消除多重共线性，除了相关预测变量的数量和关联性之外，主要取决于错误出现在研究过程的哪个地方：理论构建、具体实施或者数据搜集。“如果发现了多重共线性，具体怎么处理更像是一门艺术，而不是科学”（Menard, 2001, 80；也可参见 Pedhazur, 1982², 247）。
11. 样本量个案。对于因变量的每个变量类别，应至少是 $N=25$ 。进入模型的预测变量越多，或者模型幂次越大，则需要的个案越多。Hosmer & Lemeshow（2000, 339-347）提出了一个公式，除了样本量之外还能给定模型的幂次和检验方向。如果多个预测变量等级的组合可能导致产生很多空白单元格，则既可以从模型中剔除不重要的预测变量，也可以将预测变量等级合并起来。为了保证数据完整性，应小心地将预测变量等级合并起来。明显很高的参数预测变量或者标准误差，以及理想分割的分类图（如完全分离或准完全分离）既有可能表明模型是理想的，也可能表明存在数据问题或者模型设定错误，对此应予以检验。

可以通过简单的检验来检查预测变量具有完全分离还是准完全分离。例如，分类预测变量可以与（定类）因变量对进行交叉列表。如果在占满的单元格之间不存在值得一

提的频率差异,则分离效应是不可行的。如果只存在频率差异很大的单元格,则分离效应的概率较大。但是,如果存在这样一种配置,即只有在对角线的单元格含有频率,则很可能具有完全分离或准完全分离。例如,可以根据(定类)因变量以成组条形图的形式输出定距预测变量。根据归属于定类因变量的不同变量类别,数值具有不同的颜色。各个分布形状之间的彩色重叠区越小,分离效应的概率越大。如果有分离效应,则应不仅针对统计量,而且还要针对理论的建模以及数据搜集进行检验。SPSS的这个提示信息是正确的,但是并不完整,因为这条提示信息只局限于将统计量当作可能的原因:“在数据中可能有准完全分离。因此既不存在最大似然估计,一些参数估计也是无止尽的。”但是,用户还应继续执行一个步骤,检验所调查的模型设定的是否正确,或者是否正确地搜集了数据。

理想模型的正面例子是,接近完美地给针对因变量多个变量类别中的某一个的归属性建模(预测)。例如,来自生物学领域的一个广为人知的例子是,根据花叶的长度和宽度(成功地)对植物分类建模。

模型设定错误的反面例子可能是因果性建模不充分。例如,混淆了因变量和自变量。将针对因变量多个变量类别中的某一个的归属性“完美”地建模(预测),这样一个错误例子可能是由以下原因造成的:例如,对于预测而言,是否有人年龄太小(例如 <5 岁)或者太大(例如 >50 岁),是否使用了诸如体重和身高等变量。错误之外是,“预测变量”体重和身高实质上就是因变量,而不是预测变量。其他原因可能是在结果中的数据搜集错误,如乖离率、选择错误或者测量错误。

12. 每个参数的事件。仅有足够数量的个案是不够的,应保证目标事件(因变量)的变量类别,尤其是在有很多协变量时应以足够大的频率出现。作为经验法则, Hosmer & Lemeshow (2000, 346-347) 建议针对有些对称的分布采用至少 $N = 10 * n$ 个协变量(对于目标事件的不对称变量类别,情况更为复杂)。结果可能是只针对感兴趣的目标事件抽取个案、普遍地升高个案数量 N 以及(或者)减少协变量数量。
13. 残差(误差)。NOMREG 中的皮尔逊残差是基于成组数据的。对于逐个个案的皮尔逊残差或者其他残差,就如通常高要求的残差分析一样,应换为采用 SPSS 过程命令 LOGISTIC REGRESSION。相互独立地对所调查的特征进行取样。因此,残差应是相互独立的。残差围绕着 0 随机散布,具有恒定的方差(同方差性),呈现多项分布(只有在样本很大时才呈现正态分布),既不相关,也不与预测变量相关。残差的分布取决于样本量,对于小的样本,如果违背这个假设则被认为比大的样本要严重得多(前提条件是中央极限定理许可)。关于误差的独立性,无法保证消除自相关。对于可能通过“时间”因子而相互关联的变量,例如,在重复测量设计时,可以参见 Hosmer & Lemeshow (2000, 第 8.3 节) 著作中阐述的逻辑方法的特定前提条件和用途。
14. 模型设定。模型设定应是通过关于内容的统计学标准,而不是通过关于形式的算法来推导的。各个预测变量应相互不相关。很多作者明确建议不要使用自动变量选择的方法,但是在有保留的情况下,作为一种探索性方法也是可以使用的。对于这两种处理方法,建议如下处理:首先让内容上的有关预测变量进入模型,然后通过显著性检验

剔除统计学上的无关变量。如果模型含有预测变量之间显著的交互作用，则影响达不到显著性的预测变量也保留在模型中。

15. 例如，逐步法是根据形式上的标准（统计学上的关联）进行工作的，不适用于理论推导的建模，因为逐步法选择了内容上没有关联的预测变量。应根据可信的、关于内容的标准，对纯粹探索性的或者预测性的工作方法进行交互检验。后退法应优先于前进法，因为后退法与前进法相反，是从检验一阶交互作用开始的，因此不存在仓促剔除潜在的抑制变量的风险。但是，逐步法不会消除多重共线性，因此至少要通过交叉验证予以保障。
16. 离群值和高杠杆值。在 **NOMREG** 中，离群值诊断显然不是最佳的。建议将多元因变量分解为多个二元尺度变量，并利用 **LOGISTIC REGRESSION** 计算多个单独的离群值诊断。详细说明参见有关章节。
17. 模型拟合优度（错误分类）。通过最佳的估计，模型应可以正确地重现大部分观察到的事件。但是如果测定的回归方程对数据的拟合不佳，则可能发生个别个案实际上属于某个组，但是从输出结果来看却属于另一个组的现象。例如，在分类图中（也可参见下文关于离群值的论述）。每次在判断个案归属于哪个组时，数值低于 80% 是不可接受的，根据应用领域不同，甚至可以提出更高的要求。应检验错误分类的个案和成本（假阴性的成本通常高于假阳性的成本），必要时相应地调节分类阶段。所观察的命中率是否超过随机水平，例如，可以利用二元因变量的二项检验进行检查（参见 Bortz, 1993, 579）。
18. 预测质量检验（排除过度拟合）。一个模型在完成参数化之后，应检验其预测优度是否具有实际关联性。从而排除模型将误差增多的可能性。如果利用创建模型所基于的样本（称为“训练数据”）对模型进行检验，则命中率可能估计过高（过度拟合）。尤其是在特殊的模型中会出现过度拟合现象。原因通常是训练数据集（乖离率，分布等）的特殊性。因此，应根据经验数据，始终通过交叉验证来检验模型中是否存在过度拟合。交叉验证是预留一个或者多个其他（子）个案来进行检验（称为“验证数据”）的模型检验。在 **NOMREG** 中，可以通过将初始数据集分解为两个子集来进行交叉验证。利用第一个数据子集建立模型，利用第二个数据子集进行验证。如果一个模型表现出很大的性能差异，例如，用训练数据可以将 80% 的数据正确分类，但是利用经验数据时这个比例可能只能达到 50%，则就存在过度拟合。相反的现象就称为拟合不足，也就是忽视了真实的数据现象。拟合不足现象主要发生于太简单的模型中。
19. 在解释回归系数和胜率 **Exp(B)** 时的特殊性。（a）预测变量的尺度水平：对于胜率和回归系数的解释，其区别在于分类预测变量和定量预测变量。对于定量变量，可以用整个统一定义域的一个公共值来表达其影响；对于分类变量，则测定 $n-1$ 个变量类别或单位的数值。需要注意的是，编码会对胜率、回归系数的大小或正负号起作用（例如，二元因变量的系数可能正负号颠倒；对此参见关于参考类别的注释）。（b）分类预测变量的编码：预测变量的编码对回归系数的解释及其计算有影响。如果对于病例个案或者事件个案，编码偏离 1，并且对于对照个案，编码偏离 0（对此参见关于参考类别的备注），则必须用另外的方法测定参数。（c）非标准化回归系数与标准化回归系数：非标准化回归系数可能与标准化回归系数有很大区别，并且完全错误地

反映了各自预测变量的影响。Menard（2001）建议对于分类变量和带有自然单位的变量采用非标准化回归系数或胜率，对于没有共同单位的定比数据采用标准化回归系数。在分析之前通过将预测变量本身标准化，就可以针对定量预测变量的模型提取出回归系数，然后就可以将提取出的回归系数解释为标准化回归系数。对于带有分类预测变量的模型，处理起来会更为复杂（参见 Menard，2001）。

在线性回归中，通常建议将标准化回归系数用于比较在一个样本/总体内部的定量变量，或者用于没有共同单位的定量变量，对于后者应考虑到，其测定可能是取决于所选择的样本，并且根据模型拟合优度不同，只能有所保留地将这种测定结果普遍化。非标准化回归系数建议用于比较样本/总体之间的定量变量，或者用于具有自然/共同单位的定量变量。根据这两种回归系数的优点和缺点，Pedhazur（1982²，247-251）建议给定两种量度。如果在分析之前将数据 z 标准化，则将 β 值给定为 B 值。

20. 协变量模式的数量。具有多个变量类别的预测变量越多，则协变量模式以及由此所需的个案数量就越大。多项逻辑回归的出发点是，所有的单元格都已占满。空白、未占满的或者占满很不对称的单元格至少会导致在解释基于卡方的统计量时出现问题。如果个案数量超过协变量模式的数量，则应始终成组地测定模型参数。例如，通过皮尔逊或者 NOMREG 中的偏差。因此，作为拟合优度量度的误差或者皮尔逊不适合带有定比预测变量的模型。如果协方差模式的数量大致等于个案的数量，则应始终逐个案地测定模型参数（例如，通过成组的 Hosmer-Lemeshow 检验，只有在 SPSS 过程命令 LOGISTIC REGRESSION 中才可以使用，因此，多项模型必须分解为二项模型）。Hosmer-Lemeshow 检验是一种经过改良的皮尔逊卡方拟合优度检验，是基于分布在 10 个同样大小的组中的期望概率。对于个案数据，Hosmer-Lemeshow 检验应始终优先于误差或者皮尔逊这两个拟合优度量度。

3.5 本章所介绍的各种回归方法的比较

	用于分类回归方法的 SPSS 过程		
主要区别	LOGISTIC REGRESSION	PLUM	NOMREG
方法名称	二元逻辑回归	有序回归	多项逻辑回归
因变量类别	2 个类别	2 个类别和更多水平	2 个类别和更多水平
因变量尺度	定类	定序	定类
方法	直接和步进，预测变量的顺序可变，个案数据	直接，预测变量顺序固定	直接和步进，预测变量的顺序固定，个案数据
模型和特定假设	逻辑回归模型	累积 Logit 模型（平行性检验）	多项逻辑回归模型
参数估计值指的是...	最低的等级；通常为等级“0”（预设置），此时“1”视为“事件”	位置和阈值	最后（最高）的等级（预设置）；必要时需要重新编码
分类表	是	否	是

续表

	用于分类回归方法的 SPSS 过程		
主要区别	LOGISTIC REGRESSION	PLUM	NOMREG
模型诊断	好	差	差
保存的统计量	范围广（很多残差类型， 预测等级，Cook，dfbeta， 离差，uvam）	很少：例如皮尔逊， 残差，预测等级， β 变化	皮尔逊残差：从 SPSS 13 版开始有 AIC 和 BIC （Akaike 信息准则和 Bayes 信息准则）
建议 ^a	二元因变量，定量预测变 量，模型检验	有序因变量	绝大多数或者完全是分类 预测变量，（例如，巢式 病例对照研究）

a. 前提条件是遵守了模型假设。

第 4 章 生存分析

第 4 章介绍了生存分析的方法（寿命表，Kaplan-Meier 估计，Cox 回归）。原则上，生存分析调查的是到发生特定目标事件为止的时间。目标事件既可以是期望事件（如延长订单、受聘、学习成功、治愈等），也可以是不良事件（如被解雇、故障、旧病复发、死亡等）。这些方法有多种多样的名称，例如，Survival analyse、生存分析、time to effect 或者事件分析等，来自对目标事件的不同评估。根据对目标事件的评估不同，对于输出的图应予以不同的解释。

第 4.1 节首先介绍了生存分析的基本原则，然后介绍了面临的一些典型问题和生存分析的目标。

第 4.2 节阐述了对不同生存函数（主要包括累积生存函数、1 减去生存函数 $1-S(t)$ 、密度函数 $f(t)$ 、对数生存函数 $l(t)$ 以及危险函数 $h(t)$ ）的规定。

第 4.3 节介绍了数据截尾的入门知识。在进行生存分析时可能出现这种情况：在某些个案中，目标事件没有像期望的那样发生，也就是说，目标事件完全没有或者没有按期望的（是设定的）原因而发生。为了将这些个案与带有期望事件的个案区分开，就需要借助于截尾将其标出。本节介绍了左截尾、右截尾和区间截尾，并且解释了在（非）试验性调查设计中的截尾。

第 4.4 节以保险精算方法（寿命表法）和 Kaplan-Meier 法为例，阐述了如何用这些方法测算生存函数，以及在这个过程中如何处理截尾的个案。从第 4.6 节开始介绍 SPSS 示例。

第 4.5 节介绍了对各组进行比较的不同检验方法：对数极差检验（又称时序检验或 Mantel-Cox 检验）、Breslow 检验（又称修正的 Wilcoxon 检验，Wilcoxon 秩和检验）、Tarone-Ware 检验和似然比检验。此外，本书还归纳了一个比较性综述，以及对于解释这些检验的建议方案。

在第 4.6 节中，应用并且解释了如何利用 SPSS 的保险精算法（SPSS 过程命令 SURVIVAL）和 Kaplan-Meier 法（SPSS 过程命令 KM）。本章还阐述了 Cox 回归。针对寿命表法提出了带有或者没有因子的几个实例。在对 Kaplan-Meier 法的阐述中，介绍了带有/没有因子、带有分层变量并且针对测定置信区间的一些实例。

第 4.7 节首先介绍了 Cox 模型的特点（SPSS 过程命令 COXREG），然后将这种方法与寿命表法、Kaplan-Meier 法和线性回归进行比较。然后计算和解释了利用 SPSS 进行 Cox 回归的几种变体形式（时间独立协变量、时间相依协变量、交互作用和“模式”）。接下来的几段介绍了检验 Cox 回归的特定前提条件（主要是对截尾、多重共线性和比例性假设的分析）的方法以及如何建立对比（“偏差”、“简单”、“Helmert”等）。最后几段分别归纳了所介绍方法的各種前提条件，以及对其进行检验的方法。

4.1 生存分析概述

生存分析（也称为生存时间分析、survival analysis、寿命分析，lifetime analysis、失效时间分析、时间-效应分析、时间事件分析等）属于与时间相关的分析方法。生存分析调查了两个时间点之间间隔时间的分布和规模。第一个时间点通常定义为一次调查研究的开始时间，第二个时间点定义为某个预设事件的发生时间。因此，生存分析一方面调查了要持续多长时间才能发生某个事件，另一方面调查了是否真正发生了预期事件。生存分析的目的在于，描述生存随时间的演变过程和生存的概率。

备注：“生存分析（生存分析）”这个名称听起来是有些悲伤的，因为它表明这些分析方法的应用总是与不可避免的死亡有关，或者至少是有消极含义的（也可参见“寿命表”概念）。很多教材在介绍这些分析方法时以经历手术或者疾病的患者生存状况为例，从而也助长了这个趋势。其实仔细观察就会发现，生存分析调查的仅仅是直至发生某个指定目标事件的时间（因此也被称为时间-效应分析或者时间事件分析），这个目标事件不一定就等同于实际生活中的死亡事件或者最终事件。例如，在医学中可以调查一种药物或者疗法多久会显现（期望的）疗效。在农业领域，可以用来调查球茎什么时候开花，也可以将嫩芽凸出地表定义为目标事件。例如，只要最后凸出地表的球茎达到目标事件或者达到适当的时间界限，这个生存分析就停止。英语中对它的定义就更为中性（如“time to effect / event analysis”或者“life table”，参见 Kalbfleisch & Prentice, 2002）。下面主要使用生存分析这个表述。

关于生存分析的例子提出了下列几个问题（参见 Klein & Moeschberger 的著作，2003，第 1 章，主要是医学问题的第 1~20 页）。

- 一种药物或者疗法多长时间能显现出（期望的）疗效？
- 一名患者在接受器官移植手术（如心脏、肝脏、肾脏）后能存活多长时间？
- 吸烟者在吸完“上一根”香烟之后隔多长时间再吸烟？
- 客户对某个品牌或者供应商的忠诚度能持续多长时间？
- 客户在供应商的商店（或者网上商店）里逗留多长时间才会决定购买？
- 某个产品（汽车、手机、电灯泡等）使用多长时间后会出现故障？
- 一名驾驶新手在无故障行驶多长时间后会发第一起登记在案的交通事故？

- 与其他鸟类相比，某种鸟对栖息地的忠诚度会持续多长时间？
- 与在受到污染的环境中相比，鱼类在未受污染的环境中可以生存多长时间？
- 某个作物的种子在特定（如气候）条件下多长时间会开花？
- 一个刑事罪犯多长时间后会再次作案？

生存分析的三个目的如下（参见 Kleinbaum & Klein 的著作，2005，第 15 页）。

1. 估计和解释生存函数和/或危险函数。
例如，生存时间是立即急剧地还是逐渐缓慢地下降？
2. 比较生存函数和/或危险函数。
例如，采用某种治疗（服药，训练等）后的生存时间是否比没有采用治疗的生存时间长？
3. 调查协变量（因子）对生存时间的预测。
是否还有其他解释因子（年龄、血压等），其各自的预测有多大？

纵览和比较

标准	寿命表	Kaplan–Meier	Cox 回归
目的	描述，分析	描述，分析	描述，分析，预测
根据	观察	观察	函数值
单位	时间间隔	单个数值	单个数值
状态变量类型	事件（EVENT）	事件（EVENT）， 数据丢失（LOST）	事件（EVENT）， 数据丢失（LOST）
定量协变量	否 （超出分类的范围）	否 （超出分类的范围）	是
时间相依 协变量	否	否	是（Cox 模型 2）
协变量 （预设置）	2 个分类协变量	2 个分类协变量 （1×因子，1×层次）	N 个分类协变量， N 个定量协变量，
交互作用建模	通过菜单：有限制。通 过语句更为灵活	通过菜单：有限制。 通过语句更为灵活	通过菜单：灵活
变量选择	用户	用户	用户/自动
图 （预设置）	是	是	是（Cox 模型 1）， 否（Cox 模型 2）
截尾的图形显示	否	是	否
关于因子等级的成对检验	是	是	否
关于分层变量的成对检验	否	是	否
数据量	中等到大	小到中等	到很大
模型诊断	通过菜单：无	通过菜单：有限	通过菜单：良好
与分析有关的特殊性	可以分析汇总数据，可 以不一样大小的区间 （分别通过语句实现）	-	纳入时间相依协变量， 可以进行残差分析

4.2 生存分析的基本原理

生存分析的第一个步骤是估计生存时间（survival times）。生存函数（survivor function，也可称作 survival distribution function，SDF，survival function） $S(t)$ 用于估计所感兴趣的总体的生存时间（lifetime）。

4.2.1 生存函数 $S(t)$

对 t 进行估计的生存函数（SDF）表明，总体的一个元素的生存时间（life time） T 有多大概率将超过 t 。

$$S(t) = \Pr(T > t)$$

$S(t)$ 表示生存函数， t 表示某个时间点， T 代表某个随机选择的因素的生存时间（lifetime）。因此， T 是随机变量。

换言之， $S(t)$ 表明了生存时间 T 有多大的概率可以超过具体的时间点 t 。因此， $S(t)$ 描述了（至少）在时间点 t 后仍然生存的概率。由于在分析开始时（ $t=0$ ）所感兴趣的所有元素还都存在，因此超过这个“零”时间点仍生存的概率为 $S(t)=1$ 。随着时间的流逝（也就是时间单位 t 的数量逐渐增加），超过各个时间点仍生存的概率逐渐接近于零，也就是 $S(t)=0$ 。换言之， t 和 $S(t)$ 是反向变化的。经过的时间越长，发生某个特定事件的概率越大。作为在 $t=0$ 时的函数， $S(t)$ 从数值 1 开始，随着时间（ $t=\max$ ）的延续逐渐接近于数值 0（在大多数情况下达不到 0，因为在某个特定时间点就结束了观察）。

从图形上，可以把 $S(t)$ 想象成是从 1 朝向 0 逐渐向下的台阶，也就是单调递减的，其中各级台阶可能具有不同的陡度或宽度（也可参见下文的示例）。台阶的陡度是由正在失效的元素的数量造成的，台阶的宽度是由所经过时间点 t 的数量造成。这样的台阶越宽、越平缓，则元素的生存概率就越大。

尤其是对于比较性问题， $S(t)$ 能够提供大量的信息。例如在对两个组进行比较，调查 $S(t)$ 以多快的速度趋近于 0 时。

示例

有下列数据：

```
ID SEE T
01 1 534
02 1 463
03 2 157
04 2 98
```

其中 T 代表一种鱼的个体分别在没有受到污染和可能受到污染的海洋里的生存时间。例如，代号 03 的鱼在 2 号海洋里能生存 157 天。很明显可以看出，在 1 号海洋里的生存时间要比在 2 号海洋里长得多。对于所有的鱼，在 $t=0$ 时的生存概率 $S(t)$ 等于 1；然后在经过 $t=534$ 后，所有鱼的生存概率 $S(t)$ 等于 0。因此，鱼在 1 号海洋里的生存概率 $S(t)$ 可能大于在 2 号海洋里的生

存概率。这个例子摘录自 Schendera（2004）的著作。

4.2.2 确定生存函数 S(t)

通过公式 $S(t) = 1 - d/n$ 确定生存函数 $S(t)$ 。数值 1 指的是初始概率， d 指的是到时间点 t 时失效的元素的数量， n 指的是剩余元素的数量。

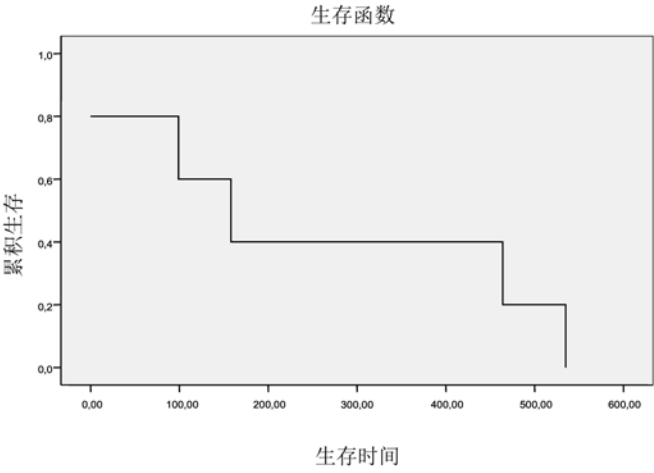
如果将这个公式套用到鱼类这个例子的数据上（为了简化内容，在这里的处理就好像是所有的鱼都是来自于同一个海洋），可以测定出下列生存概率。

t	d	n	生存函数
0	0	4	1
99	1	4	0.75
158	2	4	0.5
464	3	4	0.25
535	4	4	0

备注：为了避免一开始就产生误解， t 每增加一个数值，就称为初始数据中的某个元素失效后经过一天。

测定的 $S(t)$ 值构成从 1 开始的台阶，在每次发生一个事件时向 0 方向递减一级。如果有多个元素同时失效，则这一级台阶向下延伸的坡度比较陡（例外：最后一级台阶倾向于 0）。在出现最后一个值（ $t=535$ ）时，生存概率达到 0。对于超过这个值的生存时间，则无法做出任何判断。不允许在没有基本数据的情况下做出判断。

函数：累积生存
方法：Kaplan-Meier



解释：“累积生存函数”图展示了在一个线性尺度上的累积生存函数（Y 轴），X 轴代表生存时间。

每条垂直线代表一个失效时间点。由于这个图含有四条垂直线，因此就表明是基于上面的示例数据。这些水平线越宽，从此时到下一次失效的时间就越长。从这个图也可以看出，在第二次和第三次失效之间的时间间隔比所有其他失效之间的间隔都要长。

生存函数反映了一个个案有多大概率可以在设定时间点 t 之后仍生存（Kleinbaum & Klein, 2005, 9）。与生存函数相比，下文（第 4.2.3 节）介绍的危险函数 $h(t)$ 具有完全相反的视角。

总览表

图/SPSS 过程	寿命表 SURVIVAL	Kaplan–Meier KM	Cox–Modell COXREG
DENSITY	是	—	—
HAZARD	是	是	是
LML	—	—	是*
LOGSURV	是	是	—
OMS	是	是	—
SURVIVAL	是	是	是

* 在第 4.7.6 节预设置。

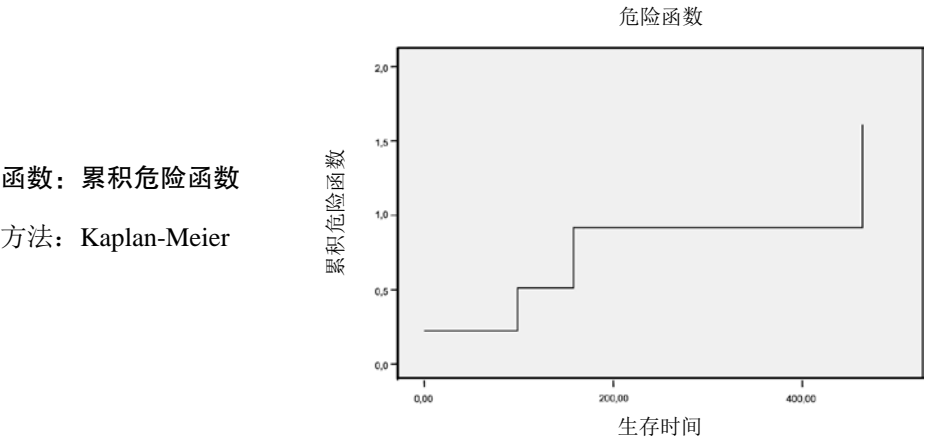
通常，生存函数是十分适合用于分析生存数据的。但是，有时也会采用其他函数。下面介绍的危险函数与生存函数相比，有一个很实际的优点：生存函数是一个超越时间范围的累积函数；相反，危险函数描述了在具体时间点或者间隔一定时间的直接效应。

4.2.3 其他函数

危险函数 $h(t)$

危险函数 $h(t)$ 反映了在一个给定时间单位内发生目标事件的直接潜力，前提条件是这个个案在直到给定时间点 t 时仍然生存。

因此，危险函数的着眼点是目标事件的直接发生，也就是某个个案的死亡。相反，生存函数 $S(t)$ 的着眼点是生存，也就是目标事件不发生。



解释：“危险函数”图展示了在一个线性尺度上的累积风险率（ Y 轴）， X 轴代表生存时间。在风险率恒定时，危险函数平行于 X 轴。如果有这样一个图，则可以认为生存时间呈指数分布。

因此，“危险函数”图表明哪些特定时间点比其他时间点“更危险”。“危险”这个词只适用于负面意义的目标事件，例如，死亡；对于正面意义的事件，“潜力更大”这个词更加适合。

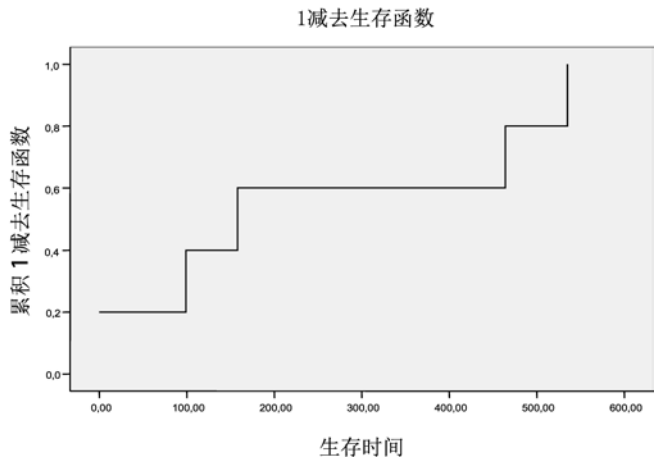
风险率总是非负的，也就是大于等于 0。 $h(t)$ 越大，发生目标事件的比率也就越高。风险率

基本上表达了一个有条件的失效比率，但是不表达概率（Kleinbaum & Klein, 2005, 10-12）。

1 减去生存函数

1 减去生存函数基于测定的 $S(t)$ 与 1 的差值，从图形上来看基本就是累积生存函数（第 4.2.2 节）的对立面。众所周知，累积生存函数描述了生存概率。相反，累积 1 减去生存函数反映了失效概率。

函数：1 减去累积生存
方法：Kaplan-Meier



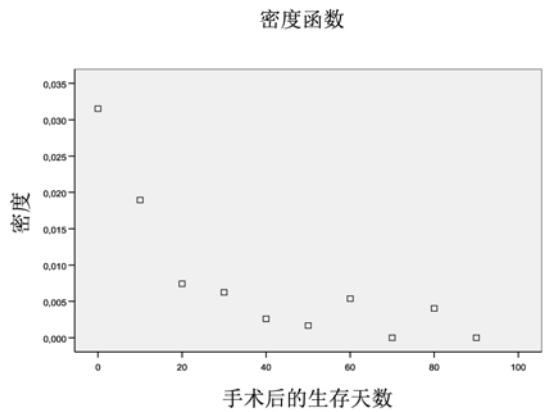
解释：“累积 1 减去生存函数”图展示了在一个线性尺度上的累积 1 减去生存函数（Y 轴）的数值。X 轴代表生存时间。累积生存函数图（第 4.2.2 节）展示了从 0.8 递减到 0 的生存概率。相反，“累积 1 减去生存函数”图（见上文）展示了从 0.2 递增到 1 的失效概率。

密度函数 $f(t)$

密度函数 $f(t)$ 是另一种描述生存时间的形式。密度函数在这方面的效力与生存函数和危险函数没有什么不同（Hosmer & Lemeshow, 1999, 11-13, 94）。

密度函数是用数学方法根据生存时间的分布函数推导得出的。这样，生存时间就与生存概率联系起来。因此，推导出的密度值表现了一种“概率”，也就是一个元素有多大概率可以“生存”到时间点 t 。

函数：密度
方法：寿命表

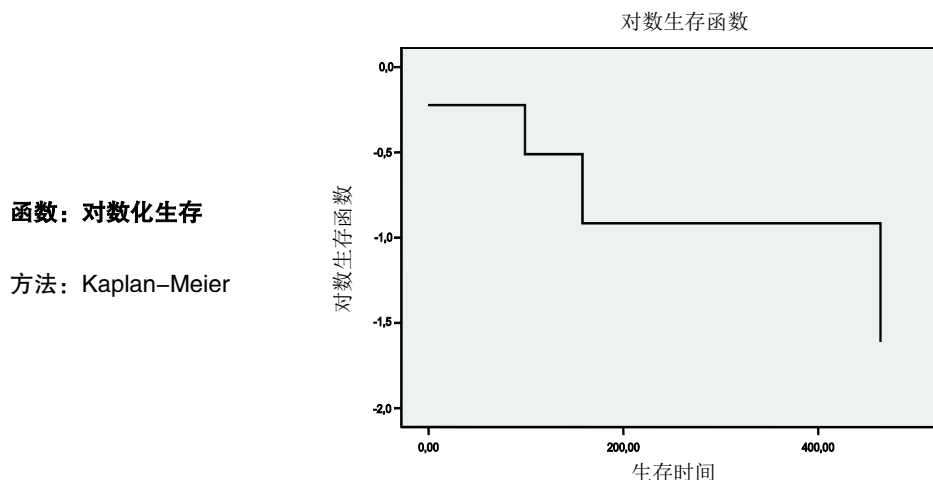


解释：“密度函数”图展示了在一个线性尺度上的累积密度值（Y 轴）。X 轴代表生存时间。

截取的各点反映了一种“概率”，也就是元素有多大概率可以“生存”到时间点 t 。

对数化生存函数 $l(t)$

对数化生存函数 $l(t)$ 是另一种描述生存时间的形式。



解释：“对数生存函数”图展示了在一个对数尺度上的累积生存函数（Y 轴）。X 轴代表生存时间。

尤其是在进行较大规模的调查研究时，不能苛求在每个时刻都可以查看初始数据中还有多少元素，或者哪些元素失效了。因此就提出了这个问题：如果目标事件没有按照计划发生，对于此时的数值应如何处理？这些数值也要纳入评估范围吗？如果是的话，该怎样操作？

4.3 截尾数据

生存分析的特征是，目标事件（也就是一定数量的元素中的某个元素，在一定的时间区间内可能失效）可能不像期望的那样发生。

其可能的主要原因是：调查研究的时间限制、退出调查研究、调查研究单位迁出或者流失。此外，也有可能是目标事件确实发生，但却不是由于所感兴趣的原因。例如，在一项癌症调查研究中，如果将死亡时间点作为目标事件，但是一名患者是由于交通事故而死亡，那么这名患者就必须作为截尾退出调查研究，否则就可能歪曲调查结果。

4.3.1 非期望事件或者未发生目标事件

现在，如果目标事件由于上述原因没有如期望的一样发生，则根据相关对象的现有数据，只能知道生存时间超过了某个数值，但是不知道超过的幅度有多大。因此，如果不知道一个个体的准确生存时间，就需要进行截尾（英语“censoring”）（Kleinbaum & Klein, 2005, 5）。本来用于确定生存函数 $S(t)$ 的准确生存时间也就仍然处于未知状态。

如果对于特定对象或者观察值，由于上述原因无法准确地确定生存时间，则将这样的观察

称为“截尾观察”，或者简单地称之为截尾。“截尾”是一种标记方式，表明了对于某个特定对象没有发生期望的目标事件（由于意外原因或者与调查研究无关的原因，见上文）。截尾分为左截尾和右截尾两种形式，左截尾是指在设定的目标事件发生之前观察对象退出调查研究，右截尾是指调查研究结束时观察对象的目标事件还未发生。还有一种是区间截尾，是上述两种形式的组合，也就是知道事件发生的时间区间，但是不知道准确的时间点。

截尾数据使得在这里运用传统的方法（例如 T 检验或者回归分析）是十分困难甚至是完全不可行的。由于这些方法无法区分正常的失效或者截尾，因此在运用时就会发生信息丢失的现象。

示例

有一项调查研究，其时间为一年（365 天），研究对象是一种鱼的个体生存时间。一只鱼（ID=02）在这个研究结束后仍然生存。而另一只鱼（ID=05）在调查研究的第 265 天到 266 天时毫无征兆地消失，再也没有找到。

ID	T	ZENSUR
01	214	1
02	365	0
03	183	1
04	257	1
05	265	0

在数据集中，这两只鱼（ID=02 和 ID=05）分别用 0 截尾（称之为 ZENSUR 变量）。ID=02 的鱼被赋予 T 值的最大值，因为这条鱼在调查研究结束后还存活（例如，在这里是第 365 天）。由于在第 265 天还看到了鱼（ID=05），但是无法证明由于期望的或者可能的原因退出这项调查研究，因此这条鱼被赋予 T 值 265。

例如，在 SPSS 文件“Breast cancer survival.sav”或者“AML survival.sav”中，将定期发生目标事件（死亡，旧病复发）的个案在变量 STATUS 中定义为“1”（“died”或“relapsed”），没有定期发生目标事件的个案编码为“0”=截尾（“censored”）。但由于这样的编码可能是随意分配的，因此强烈建议采用统一的处理方法。为了让人对这个核心数据引起一定程度的注意，下文对目标事件和截尾没有采用统一的编码（例如 4.6.1 和 4.6.2 节）。

如果命名了一个变量 ZENSUR 或者 CENSOR，并且没有将变量状态编码，而是将其对立面，也就是截尾的发生进行了编码，则在一定程度上存在混淆的风险（参见 Hosmer & Lemeshow, 1999, 2-5）。

4.3.2 对截尾数据与非截尾数据做不同处理的三个理由

为了基本可以判断出现有数据是否与期望的目标事件有关，或者如果没有关系的话，截尾数据究竟在什么情况下进行截尾处理。对截尾数据与非截尾数据进行不同的处理，还有其他原因。

第一个原因是，在截尾数据中各个元素的生存时间与期望的目标事件没有（或者至少是没有明显的）因果关系。

例如，如果用一种生物性害虫防治新方法处理了一种侵入的害虫之后，实验中所有的害虫在期望的时间内死亡，那么从这个结果就可以明确地推断出，这个方法是有效的。但是，如果有人在使用了生物性害虫防治法几天后，由于疏忽放走了实验中的害虫，就无法明确地推断出这个方法是否有效。

一方面，人们知道实验中的害虫至少到这一天仍然存活（这并不代表期望的目标事件）。但是另一方面，剩下的害虫在这几天之后已经死亡（相反，这代表期望的目标事件）。但是人们不能（足够地）确信，逃走的害虫是否就是经过这种灭虫方法之后仍然存活的害虫。

另一个重要的原因是，一个元素的抵抗力越强，其由于截尾而失效的概率，就比由于退出初始数据而失效的概率越大。

在环境毒理学领域，体现某物种特有敏感性的一个典型例子是：在扑灭一个化学品仓库的火灾而导致含有杀虫剂的消防用水流入莱茵河之后，河中的鳗鱼大多死于乙拌磷（一种有机磷）。虽然其他鱼类也遭受了灾难性的损失，但是鳗鱼对乙拌磷的反应特别敏感。

第三个原因是一种元素因其物种特有的长寿命：一种元素的寿命越长，其由于截尾而从初始数据中消失的概率，就比由于失效而消失的概率越大。

例如，如果研究一个生物群落中的不同物种，则应注意不同物种的不同寿命：一种河鳗的寿命最长可以到 85 年，鲤鱼 50 年，鲑鱼 20 年，金鱼 5 年。因此，如果一个元素失效，则应注意其物种特定的寿命。

根据这三个（和其他）原因可以得出结论：如何区分期望的目标事件和由于与调查研究无关的原因导致的数据失效，对于生存概率的说服力而言具有十分重要的意义。

4.3.3 失效数据和截尾的处理（三种方法）

在生存分析中，可以通过不同的方式处理失效数据和截尾。

方法一

将一个意外失效元素的数据从分析中完全排除，因为不知道这些数据在处理后是否仍然生存或者生存多长时间。例如，在害虫防治的那个例子中，逃走的害虫可能正是最强壮的，根据其他害虫得出的结论与其根本没有可比性。

示例

如果采用方法一，在七个星期之后出现了一个意外失效个案，则将其从分析中完全排除，从而丢失了七个星期的信息。函数就被低估。

方法二

将一个意外失效元素的生存时间纳入分析，就好像发生了原先期望的目标事件一样，前提是可以认为，初始数据中的其他元素失效的原因与其具有差不多的概率，并且这种失效完全是随机的。

例如，如果在第 157 天，也就是在目标事件发生之前还从海里钓上了一条鱼，则对这个目标事件的处理就好像它真实发生了一样。其原因是，只要钓上鱼的概率完全是随机的，那么每

条鱼的这个概率原则上是相等的。

还有一些原因在各个元素上的发生概率明显不同，并且在很大程度上是由于其他原因的系统性预测所导致的。例如，在调查研究结束时的生存者（由于个体的强壮程度和寿命不同）。这些原因就不归入这一类，而是属于方法一。在心理学或者社会学调查研究中，应注意主观的或者与个人相关的动机和理由，如迁居或者拒绝。

示例

如果在方法二中有一个七周后意外失效的个案，则将其和预期出现的个案一样纳入分析。由于没有丢失七周的信息，因此函数被估计过高，因为这个元素在被截尾后也进入了函数。

方法三

一个意外失效个案的生存时间被纳入分析，就好像是出现了一个原本期望的目标事件意义。在函数中，这个生存时间只持续到一个角色失效，然后针对失效后的那部分函数将经过截尾处理的元素剔除。

示例

方法三避免了方法一和方法二的缺点。如果有一个七周后截尾的个案，则将其纳入分析（因此没有丢失七周的信息），但是这次观察在截尾后不再作为“如预期失效”而进入函数（因此函数没有被估计过高）。

开头介绍的生存函数 $S(t)$ 公式是基于纯粹的个案数量或者失效数量，尤其是利用寿命表法（又称保险精算法）、Kaplan-Meier 法和 Cox 回归法，可以考虑到截尾的数据。

4.4 估计生存时间 $S(t)$ 的方法

最常用的截尾数据统计分析方法是保险精算法（“actuarial method”，也称为寿命表法）或者 Kaplan-Meier 法（乘积极限法）。此外，Cox 回归也可以考虑到协变量或者是与生存时间相关的变量。在第 4.4.1 节介绍了寿命表法，第 4.4.2 节介绍了 Kaplan-Meier 法，第 4.7 节介绍了 Cox 回归。Kaplan、Meier（1958）和 Cox（1972）发表的奠基文献，属于近些年来被引用次数最多的论文（参见 Ryan & Woodall, 2005, 463-464）。

4.4.1 保险精算法和寿命表法

之所以叫保险精算法，是因为这种方法是在保险精算领域发展起来的，但是也可以有效地应用于其他领域。保险精算法的基本理念是，将观察期分解为较小的、恒定的时间区间。

将所有至少在其时间区间内全程被观察到的元素纳入分析，以便估计目标事件在单个时间区间内的概率。然后，根据对所有时间区间的估计概率，计算出一个事件在不同时间点的总概率。

例如，如果将总区间 $[0, t_{\max}]$ 分解为较小的区间 $[0, t_1]$ 、 $[t_1, t_2]$ 、 $[t_2, t_3]$ 至 $[t_{n-1}, t_{\max}]$ （方括号表示各自的时间单位，“T 值”表示各自的具体极限），则通过开头介绍的公式 $S(t) = 1 - d/n$ 可以计算出生存函数 $S(t)$ 。区别在于，这个生存函数不是针对整个时间区间，而是为了

测定在各个时间区间内的概率。

按照如下方法将截尾值也考虑进来：到进行截尾时所在的时间区间为止，将截尾值算作生存元素；除此之外，假设到进行截尾时所在时间区间的中段，截尾值仍然生存。由于在每个时间区间的开始时，只计入在上一个时间区间结束时仍保留初始数据的元素，因此截尾值在下一个时间区间内不再出现。

示例

根据各个时间区间的概率计算出总时间区间的概率。在计算各个时间区间的概率时，把在一个时间区间开始时存在的元素与这个时间区间结束时仍存留的元素联系起来。在这个时间区间的中段时被截尾的元素，就称为“半”元素。

在一个时间区间（例如月）开始时有 119 个元素。在这个时间区间的过程中，5 个元素被截尾，1 个元素失效。对于这个时间区间，估计的生存概率 $S(t)$ 为：

$$S(t) = ((119 - 5 - 1) / (119 - (5/2))) = 113/116.5 = 0.97$$

最后，根据所有单个时间区间的概率，计算出总时间区间的概率。

如果截尾是完全随机的、非系统的，那么在这个前提下，可以通过保险精算法对 $S(t)$ 做出很好的估计。这种方法尤其适用于大范围的抽样。在下面的第 4.6.5 和第 4.6.6 节中将对 SPSS 示例做介绍。

4.4.2 使用 Kaplan-Meier 法估计生存时间 $S(t)$

Kaplan-Meier 法（也称为乘积极限法或者乘积极限估计）是生存分析法的一种变体形式，适用于当调查两个事件之间的时间分布时存在截尾数据的情况。Kaplan - Meier 方法的基本原理与寿命表法有两个显著区别：Kaplan-Meier 法将观察期分为较小的、长短不同（即长度可变）的时间区间，从而使失效在一个时间区间结束时出现（或者定义了区间界限）。换言之，只要有一个元素失效，“Kaplan-Meier”过程就会在该元素的失效时间点确定一个区间界限。所有超过所定义时间区间长度的元素，都会在估计相应时间区间最终事件的发生概率时予以考虑。

Kaplan-Meier 法处理截尾步骤如下：相应时间区间内的截尾不予以考虑，仅考虑正好发生在失效时间点的截尾，就如同这些截尾使这个时间区间完全生存下来一样（事实上也的确如此）。在时间区间之间的或者正好在区间界限上的截尾，不被下一个时间区间考虑。在这里，前一个时间区间的下限被从下一个时间区间中剔除（从而确定了其上限）。其上限由下一次失效的时间点确定。

严格地说，Kaplan-Meier 法估计的是当前截尾元素的生存时间，也就是直至目标事件发生的时间。其基础是：在发生目标事件的每个时间点都有对条件概率的估计，以及构成这些条件概率的乘积极限值，从而可以随时对生存概率做出估计（所以也称作乘积极限法）。在使用寿命表法时，一个时间区间内只有一个元素，因此使用这两种方法得到的生存概率相同。

在 4.6.1 至 4.6.3 节中将介绍 Kaplan-Meier 法的 SPSS 示例。以下示例说明对于截尾的处理。

4.4.3 无截尾和有截尾的示例（方法：Kaplan–Meier）

总区间的概率是根据各个区间的概率计算出来的。单个区间并不固定，而是由各个元素的具体失效时间点确定的。

无截尾的最简单计算示例

在一个为期 20 天的调查期间，有若干元素失效，即在第 6、11、16 和 20 天不存在截尾。

ID	T
01	6
02	11
03	16
04	20

除了第一个时间区间外，其他所有时间区间均可如此描述：括号始终隔开了前一个失效值（作为极限值）。

[0,6], [6,11], [11,16], [16,20]

例如，在时间区间“[11,16]”中，将 16 定义为下限，11 则表示前一个时间区间的下限。因此在此时间区间“[11,16]”中，包括所有 11 到 16（包括 16）的数值，具体地说就是从 12 到 16。

示例

用截尾值计算 $S(t)$ （用 Kaplan–Meier 法的结果作为比较）。

T	$S(t)$	$D S(t)$	截尾	$D(KM)$	Kaplan–Meier
0	1	0	-	0	$S(0) = 1$
6	0.75	1	-	1	$S(6) = 0.75$
11	0.5	2	-	2	$S(11) = 0.5$
16	0.25	3	-	3	$S(16) = 0.25$
20	0	4	-	4	$S(20) = 0$

解释：为了清楚起见，这个表格包含基于普通 $S(t)$ （左）和 Kaplan–Meier 法（右）得出的两种结果。因为 Kaplan–Meier 法不需要考虑截尾值，所以这些结果是相同的。

有截尾的第一个计算示例（I）

在一个为期 20 天的调查期间，在第 6、11、16 和 20 天有若干元素失效。在第 8 和 18 天存在截尾（参见 ZENSUR=0）。截尾没有发生在真正失效的日期。

示例

ID	T	截尾
01	6	1
02	8	0
03	11	1
04	16	1

05 18 0
06 20 1 .

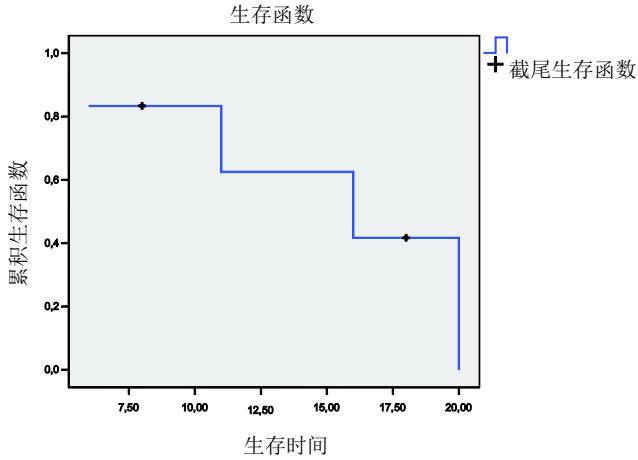
使用 Kaplan-Meier 法计算含有截尾值的 $S(t)$ (例 I)。

T	$S(t)^*$	D	截尾	$D(KM)$	Kaplan-Meier
0	1	0	-	0	$S(0) = 1$
6	0.833	1	-	1	$S(6) = 0.833$
8	0.666	2	+	1	-
11	0.5	3	-	2	$S(11) = 0.625$
16	0.333	4	-	3	$S(16) = 0.417$
18	0.166	5	+	3	-
20	0	6	-	4	$S(20) = 0$

解释：为了清楚起见，这个表格包含了使用普通 $S(t)$ (左) 和 Kaplan-Meier 法 (右) 的两种结果。 $S(t)$ 对失效和截尾做同样的处理 ($N=6$)。而 Kaplan-Meier 法却不一样 (见右)：虽然根据 $1-d/n$ 的商采用 Kaplan-Meier 法计算 $S(t)$ ，但是考虑到了在失效和截尾之间有区别的 N (参见 $D(KM)$)。换言之，截尾的个案不纳入 $S(t)$ 的计算。例如， $D(KM)$ 总是恒定的 (参见 $T=6$ 或 8)，但对于剩余个案则会予以考虑。但是，元素数量减少了这次失效 (在这个表格中未显示出)，也就意味着，只有当目标事件真实发生时，生存曲线才会改变其走向。

备注：该表格只包含使用 Kaplan-Meier 法得出的结果。例 I ($N=4$) 中作为基础的样本，比表格 II 少一个元素。这就使得结果不可能相同。从表格 I (如上所述) 和表格 II (见下文) 可以看出，在出现截尾的情况下，如何能够用 Kaplan-Meier 法有区别地计算出生存概率。这也能从生存函数的各个图中看出来。

方法：Kaplan-Meier
数据：示例 I



有截尾的第二个计算示例 (II)

在一个为期 20 天的调查中，在第 6、11、16 和 20 天有若干元素失效，在第 8 和 18 天出现截尾 (参见 ZENSUR=1)，截尾 8 (ID 19) 发生在真正失效 (ID 02) 的一天。Kaplan-Meier

法对于该种情况的处理，与失效和截尾总是出现在不同日期的情况是不同的（见下文，与例 I 相比较）。

示例

ID	T	截尾
01	6	1
02	8	1
19	8	0
03	11	1
04	16	1
05	18	0
06	20	1

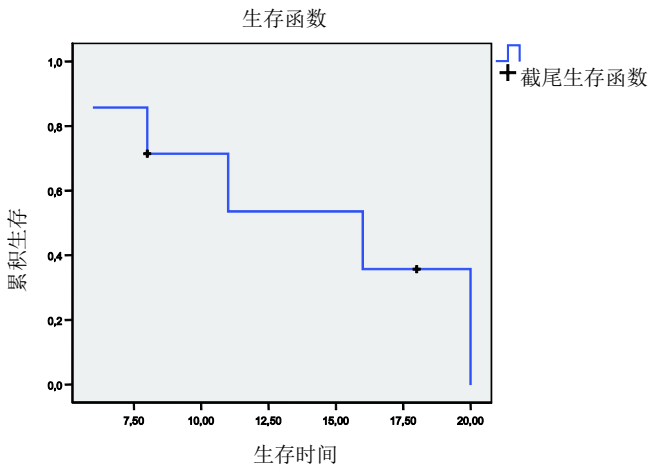
通过 Kaplan-Meier 法 II 计算含有截尾值的 $S(t)$ 。

ID	T	截尾	D(KM)	Kaplan-Meier
-	0	-	0	$S(0) = 1$
01	6	-	1	$S(6) = 0.857$
02	8	-	2	$S(8) = 0.714$
19	8	+	2	-
03	11	-	3	$S(11) = 0.536$
04	16	-	4	$S(16) = 0.357$
05	18	+	4	-
06	20	-	5	$S(20) = 0$

注：该表格只包括基于 Kaplan-Meier 法得出的结果。作为基础的样本，比上文的表格 I 多一个元素。因此，这些结果不可能与例 I 的表格内容相同。

方法：Kaplan-Meier

数据：例 II



从图中可以得出，生存曲线在时间点 8 改变其走向，而与在这个时间点发生的截尾无关，否则这个截尾就对生存曲线的走向没有影响。

到这一节为止，论述的都是一个组内的生存概率，下节将介绍从图形和统计方法上比较多个组的生存概率。

4.5 对多个组进行比较的检验

对于两个或者多个组的生存概率，可以通过图形或者统计方法进行比较。

用图形比较生存概率，就是针对每个组将一条单独的 $S(t)$ 曲线画入共同的图形中。这样就可以十分简便地对比每个组的走向。

通过不同的非参数检验，可以确保对不同组的生存概率进行比较。在这里，SPSS 提供的方法有对数秩检验（也就是 Mantel Cox 检验）、Breslow 检验（修正的 Wilcoxon 检验，也可以称作 Wilcoxon-Gehan 比分检验）、Tarone-Ware 检验和在 SPSS 16 版本中仅对 Cox 回归提供的似然比检验。SPSS 没有提供的检验方法，如 Peto Harrington 检验或者 Fleming Harrington 检验，在这里不做介绍。

用于生存分析的 SPSS 过程命令提供不同的检验方法。

TEST/SPSS 过程命令	寿命表 SURVIVAL	Kaplan–Meier KM	Cox 模型 COXREG
对数秩检验	-	是	-
Breslow 检验/ Wilcoxon-Gehan 比分检验	是	是	-
Tarone-Ware 检验	-	是	-
似然比检验	-	-	是

所有检验都是基于这个零假设：就生存时间而言，各组之间不存在差异，并且基于相应的备择假设，各组的生存概率有所区别。每种检验方法都有各自的特点，下面将做简要介绍。在这里，对数秩检验、Breslow 检验和 Tarone-Ware 检验都是基于相同的初始公式。

这里的阐述参考了 Kleinbaum & Klein （2005，57-70）、Klein & Moeschberger （2003，205-234）和 Hosmer & Lemeshow （1999，71-72）的著作。

基本公式

$$\frac{(\sum_j w(t_j)(m_{ij} - e_{ij}))^2}{\text{var}(\sum_j w(t_j)(m_{ij} - e_{ij}))}$$

其中：

- $i=1,2$
- j =第 j 个失效时间点
- $w(t_j)$ = 第 j 个失效时间点的权重

这些检验的区别只在于各个失效时间点的加权。

TEST / SPSS 过程命令	加权 $w(t_j)$
对数秩检验	1
Breslow 检验/ Wilcoxon-Gehan 比分检验	n_j
Tarone-Ware 检验	$\sqrt{n_j}$

对数秩检验

对数秩检验（又称 Mantel Cox 检验）检验的是这样一个零假设：两个或者两个以上组的生存曲线是相同的。对数秩检验对所有数值进行相同的加权，近似于卡方分布，并且体现出自由度的数量，也就是组的数量减 1。对数秩检验对所有失效进行同样的处理。

对数秩检验基于对大量抽样的卡方检验，将观察个案与期望个案进行比较，且不论个案属于哪个组。按照发生目标事件的时间顺序，将所有数据划分成不同类别。在每个发生了一个或多个目标事件的时间点，计算出这个零假设的期望值：各组中的目标事件概率相同（对于截尾个案一直考虑到其出现时，之后不再考虑）。将由此得出的卡方值，与表示自由度数量以及所选择显著性水平数量的“关键”值做比较。此时，如果这个关键值大于关键基准值，则可以拒绝零假设，并且得出如下结论：各组从统计学角度来看有显著的区别。

用途与偏误

由于各个失效的加权是同样大的，对数秩检验尤其适用于预测将发生的这种情况：各个效应平均的对某个生存个案产生影响，或者存在一个恒定风险。与其他检验相比，对数秩检验的偏误在于：对于后期事件（差异）的加权可能趋于增强，从而有可能尽管这些曲线除了末端之外的走向都是相同的，但是仍输出各条不同曲线这个结果。

Breslow 检验

Breslow 检验（又称修正的 Wilcoxon 检验、Wilcoxon Gehan 比分检验）检验的是这样一个零假设：两个或者两个以上组的生存曲线是相同的。用每个时间点有危险个案的数量 n 对各个时间点进行加权。由于 n （个案的数量）通常越来越小，所以 Wilcoxon 检验并不对所有数值进行平均处理，而是在曲线始端对数值的加权较大。失效越早出现，它的权重相应地就越大。

Breslow 检验是 Wilcoxon 秩和检验的进一步发展，不能将二者混淆。与秩和检验相反的是，修正的 Wilcoxon 检验能够把截尾值纳入考虑范围。

用途与偏误

由于 Breslow 检验在曲线前端对事件（差异）的加权比较大，因此 Breslow 检验尤其适用于这一情况：可以预测到，某种治疗对某个生存个案的效应在调查研究的开始阶段最大，之后越来越小。

Breslow 检验的偏误在于，尽管各条曲线除了前端不同之外走向都是相同的，但是仍趋向于输出各条不同曲线这个结果。如果出现这样的情况，则最好使用对数秩检验。

Tarone-Ware 检验

Tarone-Ware 检验的对象是这样一个零假设：两个或两个以上组的生存曲线是相同的。根据每个时间点危险个案数量的平方根，对各个时间点进行加权，从而使前期事件在这里有一个较大的影响，但是达不到 Breslow 检验时的程度。由于加权较小，Tarone-Ware 检验的数值大小通常介于其他两个检验之间。

用途与偏误

建议在失效既不集中在分布前端也不集中在分布末端时使用 Tarone-Ware 检验。当函数以较大的时间区间分布时，Tarone-Ware 检验尤其适用。

比较性总结

在待比较各组的曲线走向重合的情况下，不推荐使用前三种检验方法（对数秩检验、Breslow 检验和 Tarone-Ware 检验）。此外，对于曲线的解释也应该考虑一些其他特征，如样本范围、事件发生的时间点、事件数量以及截尾数量。

这三个检验方法将期望事件和观察事件的数量（包括截尾的数量）做比较，不同之处在于对各个个案的加权。对数秩检验对所有事件进行相同的加权，Breslow 检验根据个案数量对所有个案进行加权，Tarone-Ware 检验则用个案的平方根进行加权。这样得到的结果是，后两个检验方法对前期事件加权更大，而对数秩检验对后期事件的加权更大。

在理想情况下，这三个检验方法的结果应是相互支持的。这样做的好处在于，得出的结果不依赖于检验方法。如果三个检验方法的结果有所偏差，则可以考虑下文中差异化的说明。

优先采用最适合分布状况的检验方法：如果不同组的事件概率是呈相乘关系的，或者具有恒定的风险，那么最适合采用对数秩检验，这种检验方法同样比较适合后期时间点。否则应选择 Breslow 检验，但是如果（由于前期个案的强烈影响）存在很多截尾，则这种检验方法就有问题了。当生存分布呈现较大的时间区间时，Tarone-Ware 检验就体现出它的优势。需要补充的是，这些检验方法原本是用于大量抽样的，出发点是目标事件的分布不受截尾分布的影响。因此，如果只出现很少的目标事件或者抽样数量很小时，应谨慎使用这些检验方法（Klein & Moeschberger, 2003, 214）。

鉴于这个调查结果，不应在事后选择显著性最大的检验方法，而应在事前选择最适合这个问题的检验方法（Kleinbaum & Klein, 2005, 66）。例如，如果所要研究的是手术后即刻的进展情况，则应优先采用 Breslow 检验，因为该检验方法能够更好地识别初期的差异。相反，如果所感兴趣的是较晚发生的事件，那么应该采用对数秩检验。

如果分布状况事先并不明确，则事后选定的检验方法至少应支持生存图中显示的分布状况。Hosmer & Lemeshow（1999, 71-73）在著作中建议，给出所有可行的检验方法，以便让人区分出所调查的函数哪里相似，哪里不同。

似然比检验

似然比检验（又称-2Log(LR)检验、-2LL 检验）仅在 Cox 回归时提供。当相互比较的各组的曲线相互重合时，适用似然比检验。似然比检验近似于呈现卡方分布。

如果似然比检验和 Wald 检验得出不同结果，则在不确定情况下优先采纳似然比检验的结果，因为 Wald 检验的检验性能较差（参见 Kleinbaum & Klein, 2005, 90）。

4.6 利用 SPSS 进行生存分析

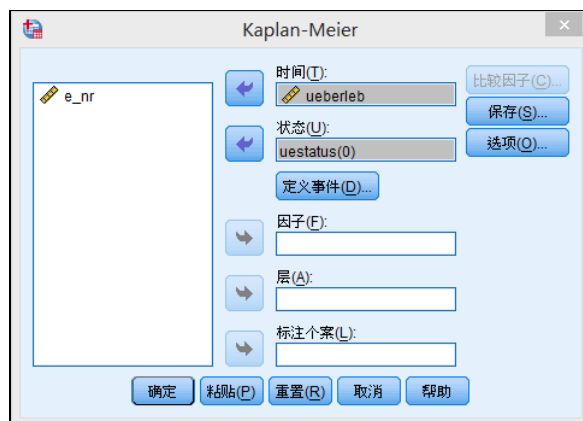
SPSS 通过过程命令 KM（Kaplan-Meier，第 4.6.1 至第 4.6.3 节）与 SURVIVAL（死亡表，第 4.6.5 节，第 4.6.6 节）来计算非参数生存分析。半参数生存分析（Cox 回归，带有时间相依协变量的 Cox）则通过过程命令 COXREG 来实现（参见第 4.7 节）。

4.6.1 示例：无因子 Kaplan-Meier 法

通过 Kaplan-Meier 分析，调查患者在术后的生存时间。本例中出现的目标事件编码为“0”。

在 SPSS 程序主界面选择以下菜单项：分析→生存函数→Kaplan-Meier...

将变量 UEBERLEB（生存）拖动至“时间”栏，将变量 UESTATUS 拖动至“状态”栏。在“定义事件”栏下面，给定所发生目标事件的编码，例如，代码“0”。单击“继续”按钮。



子窗口“比较因子”：不做任何设置。

子窗口“选项”：在“绘图”下选择“生存函数”与“危险函数”。确保“统计量”一项下的“生存表”、“生存时间平均值和中位数”以及“四分位数”已勾选。单击“继续”按钮。

子窗口“保存”：不做任何设置。

单击“确定”按钮开始计算。

语句：

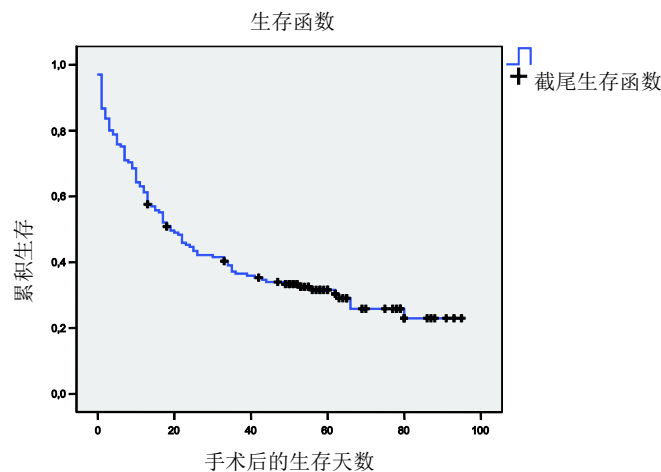
KM

```
UEBERLEB
/STATUS=UESTATUS(0)
/PRINT TABLE MEAN
/PERCENTILES
/PLOT SURVIVAL HAZARD .
```

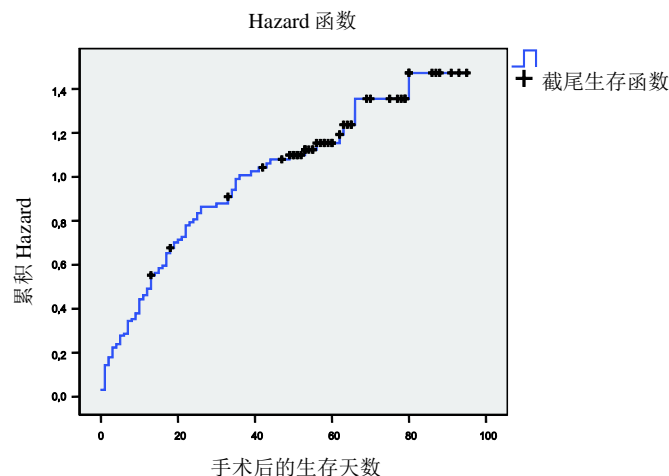
注：KM 命令调出根据 Kaplan-Meier 方法的生存分析。UEBERLEB 代表生存时间。在 STATUS=后面给出状态变量 UESTATUS。括号中的编码（“0”）代表目标事件，这个编码并不代表截尾。在 PRINT 后面，通过 TABLE 和 MEAN 调出表格“生存表”与“生存时间平均值和中位数”。PERCENTILES 命令可生成表格“百分位数”。PLOT 命令调出生存曲线图（SURVIVAL：累积生存函数，HAZARD：累积危险函数）。

输出结果

函数：累积生存
方法：Kaplan-Meier



函数：累积 Hazard
方法：Kaplan-Meier



从“累积生存函数”图可以看出，生存概率随着时间推移逐渐减小。但由于最后一个数值是截尾的（见下文的“生存表”），所以生存概率并未达到零。

对直线走向的解释取决于所调查的研究对象。如果一个效应是正向的（例如，药物疗效的出现时间点尽可能早），则累积生存函数直线在理想情况下应尽可能地靠近左边，呈现陡直下降。相反，如果一个效应是负向的（例如，由于环境因素使得死亡时间尽可能迟），则累积生存函数直线在理想情况下应尽可能缓慢地下降。

从两个图中可以看出，截尾向末端累积。对此可解释为，患者往往在简单手术之后提前出院，因而在这项调查结束之前就已经被排除。

个案处理总结

总量	事件数量	截尾	
		N	百分比
165	116	49	29.7%

表“个案处理总结”给出了个案的数量（ $N=165$ ）、出现的目标事件数量（ $N=116$ ）以及截尾数量（ $N=49$, 29.7%）。该表并非无关紧要，因为在有些情况下，“生存表”中的“状态”栏并不会列出全部截尾（参见第 4.6.2 节）。

注：下面的“生存表”的内容比较广泛，但为了节省空间，将其做了简短化处理。但是“生存表”的内容并不是要表明，Kaplan-Meier 法一定是适用于大量抽样的，有可能借助寿命表的分析更为适用。

生存表						
	时间	状态	某时间点的生存者累积比例		累积事件的 数量	剩余个案的 数量
			估计值	标准误差		
1	.000	死亡	.	.	1	164
2	.000	死亡	.	.	2	163
3	.000	死亡	.	.	3	162
4	.000	死亡	.	.	4	161
5	.000	死亡	.970	.013	5	160
6	1.000	死亡	.	.	6	159
7	1.000	死亡	.	.	7	158
8	1.000	死亡	.	.	8	157
9	1.000	死亡	.	.	9	156
10	1.000	死亡	.	.	10	155
11	1.000	死亡	.	.	11	154
12	1.000	死亡	.	.	12	153
13	1.000	死亡	.	.	13	152
14	1.000	死亡	.	.	14	151
15	1.000	死亡	.	.	15	150
		...省略...				
150	69.000	截尾	.	.	115	15
151	70.000	截尾	.	.	115	14
152	75.000	截尾	.	.	115	13
153	77.000	截尾	.	.	115	12
154	78.000	截尾	.	.	115	11
155	79.000	截尾	.	.	115	10
156	79.000	截尾	.	.	115	9
157	80.000	截尾	.229	.045	116	8
158	80.000	截尾	.	.	116	7
159	80.000	截尾	.	.	116	6
160	86.000	截尾	.	.	116	5
161	87.000	截尾	.	.	116	4
162	88.000	截尾	.	.	116	3
163	91.000	截尾	.	.	116	2
164	93.000	截尾	.	.	116	1
165	95.000	截尾	.	.	116	0

“生存表”中列举了 165 例个案，其中包括 116 个目标事件（“死亡”）和 49 例截尾（“截尾”）。“时间”一栏给出了时间点，例如，直至发生目标事件或者进行截尾时还有 0、1 或者 95 天。

“状态”一栏表明，相应的个案究竟是作为目标事件（“死亡”）出现还是被截尾（“截尾”）。第三列“某时间点的生存者累积比例”的“估计值”一项下给出了从表格开始到相应时间点（生存概率）这个阶段个案不断减少的比例以及对应的标准误差。

例如，第 157 一行应解释为：在第 157 天由于出现了一个目标事件，初始数据中的一个元素失效，此时的生存概率为 0.229。Kaplan-Meier 法估计的标准误差为 0.049。在第 157 天出现了第 116 个，即最后一个目标事件（参见“累积事件的数量”一列，截尾元素不计入在内）。此时初始数据中还剩 8 个（截尾）元素（参见“剩余个案的数量”一列，截尾元素计入在内）。

如果多个个案同时发生目标事件（“死亡”）（例如，个案 1-5 在时间点 0 时全部死亡），则估计值只给出一次（例如，个案 5 的所在行），但适用于在这个时间点发生目标事件的所有个案。

例 4.6.2 解释了 Kaplan-Meier 法如何处理截尾值。截尾个案没有被纳入 $S(t)$ 的计算，但剩余个案减去了这个失效（参见“剩余个案的数量”一列）。

生存时间平均值和中位数							
平均值 ^a				中位数			
估计值	标准误差	95%置信区间		估计值	标准误差	95%置信区间	
		下限	上限			下限	上限
37.384	2.969	31.565	43.204	19.0	2.974	13.170	24.830

a. 如果估计值是截尾的，则限制到不超过最长生存时间。

从表“生存时间平均值和中位数”中可以查取“平均生存时间”的估计值。例如，等于平均值的是估计值 37.4，此时 95%置信区间下限和上限分别为 31.6 和 43.2。中位数估计为 19.0，此时 95%置信区间下限和上限分别为 13.2 和 24.8。标准误差分别约为 3.0。

建议将中位数用作点估计值或者集中趋势测度，以对“平均生存时间”做出说明。平均值在这里有以下三个弱点。

- 如果在最后一个目标事件后还出现截尾，则平均值向下偏误。
- 如果截尾出现多次，则对分布上半部的估计往往较差，这又影响了对平均值的估计。
- 平均值容易受到离群值的影响。

从逻辑上讲，对于带有截尾的生存数据，中位数是较为适宜的测度。中位数反映了一个时间点的准确数值，这个时间点将整个分布分为生存时间高于和低于中位数的个案，并且这两种个案各占 50%（参见对百分位数解释的详述）。

百分位数					
25.0%		50.0%		75.0%	
估计值	标准误差	估计值	标准误差	估计值	标准误差
80.00	.	19.00	2.974	7.00	1.842

表格“百分位数”包括对四分位数和相应标准误差的估计值。这些数据可用于生存时间相互之间的比较。例如，第一个四分位数（25%）为 80.0（所有其他数据均在 80 处截尾，因此没有标准误差）（参见“生存表”）。第二个四分位数（50%，中位数，参见上文）为 19.0，第三个四分位数（75%）为 7.0。解释这个表格时必须注意，尽管生存时间持续增长，但百分位数的数值却逐渐减少，因此可以解释为二者呈反向（例如，第一个四分位数为 80.0 → 第二个四分位数为 19.0 → 第三个四分位数为 7.0）。

与前面列举的生存表相比，对百分位数的解释也取决于所调查的研究对象。例如，第一个四分位数（25%）表明，在时间点 80 时，初始数据中只含有大约 25% 的元素。

百分位数的比较往往能够反映很多情况。百分位数定义了分布区间，百分位数之间的差异或者百分位数的具体位置则能体现出时间区间的宽度。25% 与 50% 之间隔了 61 天，50% 与 75% 之间则隔了 12 天。这意味着，从第 7 天（75%）开始，在较短时间（12 天）内有 25% 的元素从初始数据中失效。这就导致在时间点第 19 天时只剩下了 50% 的元素。相反，从第 19 天（50%）开始，初始数据中继续有 25% 的元素相对缓慢地失效（超过 61 天），从而在时间点第 80 天时仍剩余 25% 的元素。

补充：对截尾的解释

如果此处已经完成了 Kaplan-Meier 分析，那么就忽视了一个重要结果，即一直延伸到分布末端的截尾累积。对这一现象的解释取决于对这项调查研究的设计，但是也可能取决于个人效应、治疗效应和/或设计效应这些层面（参见 Rasch 等，1996、Rasch et al., 1998, Kap. 6.31、Sarris, 1992）。

如果进行的是一个准实验，例如，相关元素不是同时纳入调查研究，或者不检验是否有任何异常（如干扰变量），则有下列解释的可能性。

设计效应：调查研究过于短暂。与其他个案不同，若较晚纳入调查研究的个案，在调查研究结束时仍存活，则必须将其截尾。

个人效应：调查研究由两个子总体组成，即比较稳健的个案和比较敏感的个案。稳健子总体经过这项调查研究存活下来，因此在结束时将其截尾（选择效应）。

即使是标准实验，也不能免受非随机效应（偏误）。恰恰相反，由于存在非随机截尾，原先的标准实验在调查研究结束时已经不再是标准实验了，因为数据不再是随机过程的产物。可能存在的干扰效应如下。

设计效应：尽管所有在调查研究结束时仍存活的个案都是同时被纳入的，但是这项调查研究对于大量个案来说仍然过短（有可能与“稳健性”不受控的偏误产生交互作用，参见第 4.7.5 节的例 I）。

治疗效应：研究计划低估了治疗效应。例如，在一项医药研究中，由于不可预测因素出现了多种意外的副作用，从而由于医学或者伦理原因将个案从这项研究中剔除，并且必须予以截尾（由于治疗特定的反应引起的偏误）。

个人效应：一组个案参加了一项医药研究，在研究期间获悉了另一种可选治疗方法具有更

好的疗效，因此自愿退出这项研究，以便接受另一种治疗（自我选择引起的偏误）。

此时的截尾透露出大量信息。截尾与时间（在调查研究结束时的累积）的关系仍需更为详细地探索，并且截尾也可能是由一个或多个仍需详细探索的因素造成的。随机截尾的前提并未满足。在调查研究过程中的因果因素和/或环境因素可能不是一成不变的。不断变化的因果因素和/或环境条件违背了这个隐含的生存分析基本假设。对截尾的解释会引出一个无法忽略的现象，对其必须做进一步的探索，但又会从根本上影响到对结果的单一因果解释。笔者在此建议做进一步的数据探索（如必要时通过适当的协变量来检验截尾），对研究设计进行回顾性分析。

4.6.2 示例：采用因子的 Kaplan-Meier 法

有三个（虚构的）湖泊，污染程度依次为：“湖泊 1”，受保护；“湖泊 2”，普通；“湖泊 3”，受污染。对这三个湖泊中的某种鱼的生存概率，分别用统计法和图形法进行研究。目标事件编码为“1”，截尾数据用“0”标识。所使用的生存分析采取 Kaplan-Meier 法。

在 SPSS 程序主界面选择以下菜单项：分析→生存函数→Kaplan-Meier...

把变量 TIME 拖动至“时间”栏，将变量 STATUS 拖动至“状态”栏。在工作栏“定义事件”下给出正在出现的目标事件的编码，例如代码“1”。把变量“湖泊”拖动至“因子”栏，单击“继续”按钮。

子窗口“比较因子”。在“检验统计量”一项下勾选 Log Rang、Breslow 与 Tarone-Ware 选项。选择选项“所有层共同”（从而检验所有因子等级的生存曲线的相等性）。确保未勾选“因子等级的线性趋势”选项。该检验仅适用于因子等级呈现自然顺序的情况（例如，如果因子等级是由不同大小、理想情况下呈现等距差值的变量构成的。）。单击“继续”按钮。

注：因子变量与分层变量之间的区别在于：因子变量可以用于直接比较生存分布，而分层变量则会输出各自的分析结果（参见第 4.6.3 节）。

子窗口“选项”。在“绘图”一项下勾选选项“生存函数”、“1 减去生存函数”、“Hazard”以及“对数生存函数”。确保“统计量”一项下的选项“生存表”，“生存时间平均值和中位数”以及“四分位数”已勾选。单击“继续”按钮。

子窗口“保存”。此例中，不要把生存分布、生存分布的标准误差、风险率以及累积事件保存为新变量。

单击“确定”按钮开始计算。

语句：

KM

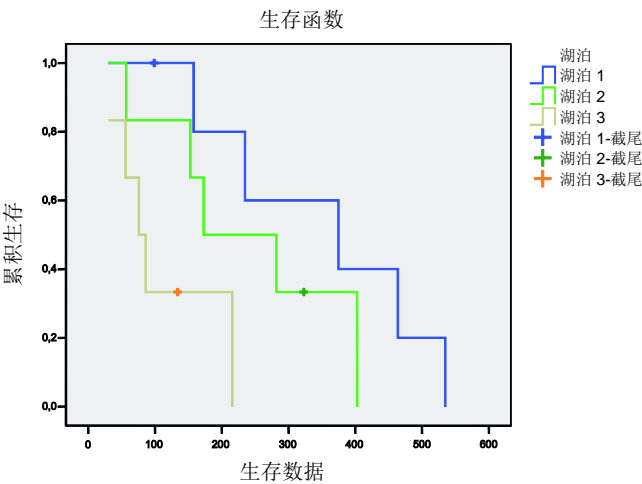
```
TIME BY SEE /STATUS=STATUS(1)
/PRINT TABLE MEAN
/PERCENTILES
/PLOT SURVIVAL OMS HAZARD LOGSURV
/TEST LOGRANK BRESLOW TARONE
/COMPARE OVERALL POOLED .
```

注：KM 命令调出利用 Kaplan-Meier 法的生存分析。TIME 代表生存时间，BY 后为分组因子“湖泊”。STATUS=的后面是状态变量 STATUS。括号中的编码（“1”）代表目标事件。截尾编码（“0”）无须单独给出。PRINT 后面，通过 TABLE 和 MEAN 调出表格“生存表”与“生存时间平均值和中位数”。通过 PERCENTILES 可以创建“百分位数”表。

如果没有给出数值从 0 到 100 的百分位数，则会默认输出四分位数。PLOT 命令可调出生存曲线图，包括 SURVIVAL（累积生存函数）、OMS（累积 1 减去生存函数）、HAZARD（累积危险函数）以及 LOGSURV（对数生存函数）。TEST 命令可通过 LOGRANK、BRESLOW 以及 TARONE 调出对组别差异的相应检验。在 COMPARE 后面确定比较方向：OVERALL POOLED 调出组间比较，即通过一次对生存曲线相等性的检验，相互比较因子等级。

输出结果

方法：Kaplan-Meier
数据：湖泊例子



生存概率在受保护的湖泊（最右边“湖泊 1”）中最高，在受污染的湖泊（最左边“湖泊 3”）中最低。普通湖泊（中间“湖泊 2”）的生存概率在两者之间。

对图中各条线的解释取决于所调查的研究对象。如果一个效应是正向的（例如，药物有利疗效出现的时间点），则分布线在理想情况下应尽可能靠左，在其他分布线下方。如果一个效应是负向的（例如，由于环境污染导致的死亡时间点），则分布线在理想情况下应尽可能靠右，在其他分布线上方。

注：如果每次将最后一个数值截尾，则相应的曲线无法达到 0（参见第 4.6.1 节）。

个案处理总结

湖泊	总数	事件数量	截尾	
			N	百分比
湖泊 1	6	5	1	16.7%
湖泊 2	6	5	1	16.7%
湖泊 3	6	5	1	16.7%
汇总	18	15	3	16.7%

表“个案处理总结”给出了个案数量、所出现的目标事件数量以及截尾数量。例如，因子等级“湖泊 1”包含 6 个个案，其中有 5 个出现了目标事件，一个出现截尾。由于后续表“生

存表”的“状态”一列并没有列出全部截尾，因此这个表格比较重要。

生存表							
湖泊	时间	状态	某时间点的生存者累积比例		累积事件的 数量	剩余个案的 数量	
			估计值	标准误差			
湖泊 1	1	158.000	出现	.800	.179	1	4
	2	235.000	出现	.600	.219	2	3
	3	375.000	出现	.400	.219	3	2
	4	464.000	出现	.200	.179	4	1
	5	535.000	出现	.000	.000	5	0
湖泊 2	1	57.000	出现	.833	.152	1	5
	2	153.000	出现	.667	.192	2	4
	3	173.000	出现	.500	.204	3	3
	4	282.000	出现	.333	.192	4	2
	5	323.000	截尾	.	.	4	1
	6	403.000	出现	.000	.000	5	0
湖泊 3	1	31.000	出现	.833	.152	1	5
	2	56.000	出现	.667	.192	2	4
	3	76.000	出现	.500	.204	3	3
	4	86.000	出现	.333	.192	4	2
	5	134.000	截尾	.	.	4	1
	6	216.000	出现	.000	.000	5	0

“生存表”代表性地展示了因子等级“湖泊 1”的输出结果。对湖泊 1 进行了 6 次观察，其中 1 次数据被截尾（相反的情况参见“湖泊 2”），5 次发生目标事件（“状态”）。即使某一行中的第一个个案同时也包括一个截尾，在“状态”一列下也不会将其列为截尾。但在表“个案处理总结”中可以看出，同“剩余个案的数量”列一样，“湖泊 1”中也出现了一次截尾，不是出现在个案 5（共 6 个），而是在个案 4（原因是截尾个案被提前）。

“时间”一列给出了时间点，例如，158、235 等，表示截止发生目标事件或者截尾的天数。“状态”一列则表明，相应的个案是作为目标事件出现（“出现”）还是被截尾（“截尾”）。第三列“某时间点的生存者累积比例”中，“估计值”栏下给出了从表格开始时间到相应时间点（生存概率）的个案逐渐减少的比例，以及对应的标准误差。

例如，时间=235.00 一行应理解为：在第 235 天，初始数据中一个元素失效，截至这个时间点有 60%的个案失效，即生存概率为 0.60。使用 Kaplan-Meier 法进行估计，对应的标准误差为 0.219。第 235 天出现第二个目标事件（参见“累积事件的数量”一列），初始数据中还剩余 3 个元素（参见“剩余个案的数量”一列）。

从表中的“湖泊 2”（时间=282）可以看出 Kaplan-Meier 法对于截尾值的处理。截尾个案不会纳入对 $S(t)$ 的计算。 $D(KM)$ 保持恒定（参见“累积事件的数量”），但剩余个案中失效个案减少一个（参见“剩余个案的数量”）。

生存时间平均值和中位数

湖泊	平均值 ^a				中位数			
	估计值	标准误差	95%置信区间		估计值	标准误差	95%置信区间	
			下限	上限			下限	上限
湖泊 1	353.400	69.956	216.286	490.514	375.000	153.362	74.140	675.590
湖泊 2	245.167	59.011	129.056	360.827	173.000	78.996	18.168	327.832
湖泊 3	113.500	33.997	46.865	180.135	76.000	18.371	39.993	112.007
总值	242.744	40.782	162.812	322.675	216.000	48.648	120.650	311.350

a. 如果估计值是截尾的，则限制到不超过最长生存时间。

从表“生存时间平均值和中位数”中能够查到均值和中位数的估计值，用以对三个湖泊的生存时间进行第一次比较。例如，湖泊 1 平均值的估计值是 353.4，此时 95%置信区间下限和上限分别为 216.286 和 490.514。湖泊 1 的中位数为 375.0，标准误差为 153.362。三个湖泊的置信区间相互重叠，例如，湖泊 2 和湖泊 3 的置信区间上限（327.8 与 112.0）就落在湖泊 1 的置信区间（74.1，675.6）内。因此，生存时间平均值出现较大差异的概率并不高。

受保护的湖泊生存时间平均值（353.4）和中位数（375.0）最高，相反，受污染的湖泊则最低（113.5，76.0）。正常湖泊的生存时间居于这两者之间。

注：其他两个湖泊的输出结果与此类似，故不再详述。置信区间较大、标准误差较高的现象是由于样本较小造成的。上面这个例子中的样本过小，不足以得出置信区间的可靠估计值。

为了说明“平均生存时间”，建议将中位数作为点估计值或者衡量主要趋势的尺度，因为中位数不受离群值的影响。中位数给出的是一个具体数值，在达到这个数值之前，50%个案的生存时间低于平均值。如果采用算术平均值，为了能够做出有效的说明，则必须列出所有个案。

百分位数

湖泊	25.0%		50.0%		75.0%	
	估计值	标准误差	估计值	标准误差	估计值	标准误差
湖泊 1	464.000	79.604	375.000	153.362	235.000	84.349
湖泊 2	403.000		173.000	78.996	153.000	110.851
湖泊 3	216.000		76.000	18.371	56.000	28.868
总共	375.000	83.921	216.000	48.648	86.000	57.335

“百分位数”表包含了四分位数及其标准误差的估计值。这些数据可用于比较三个湖泊的生存时间。例如，湖泊 1 的第一个四分位数（25%）为 464.0，相应的标准误差为 79.6。第二个四分位数（50%，中位数，参见上文）为 375.0，第三个（75%）是 235.0。对此进行解释时必须注意，尽管生存时间不断增长，但百分位数的数值却逐渐地呈降序排列，因此也可以解释为两者的趋势相反（例如，QI: 464.0，QII: 375.0，QIII: 235.0）。

与前面列出的生存时间图形类似，对百分位数位置的解释也取决于所调查的研究对象。例如，湖泊 1 的第一个四分位数（25%）表明，在时间点 464 时，初始数据中只含有大约 25%的元素。

百分位数定义了数值的分布段，百分位数之间的差值及其具体位置则说明了一个时间区间的宽度。在“湖泊 1”中，25%与 50%之间的差值为 89 天，50%与 75%之间则是 140 天，25%与 75%之间是 229 天。在“湖泊 2”中，这三个数值分别为 230 天（25%~50%）、20 天（50%~75%）与 250 天（25%~75%）。“湖泊 3”的则为 140 天（25%~50%）、20 天（50%~75%）与 160 天（25%~75%）。

（为清楚起见）如果仅限于对 25%~50%的数据进行比较，则可以清楚地看出：湖泊 1 的 25%~50%时间区间为 89 天，湖泊 2 为 230 天，湖泊 3 则为 140 天。这意味着，湖泊 1 中，25%的个案集中在相对较短的时间周期内，而在湖泊 3 中，个案则分布在较长的时间周期 140 天内。但是，时间区间的宽度与其位置不能混淆（参照百分位数截然不同的位置）。因此对于湖泊 1 可以确定，在 375（50%）与 464（25%）之间继续有 25%的元素从初始数据中失效，但是发生在较短的时间周期内（89 天）内。相反，对于湖泊 3 可以确定的是，在一个非常早的时间点（76.0，50%）时就已经有 25%的元素从初始数据中以较慢的速度（140 天）失效，直至达到 216（QI）。

总体比较

	卡方	自由度	显著性
对数秩 (Mantel-Cox)	8.536	2	0.014
Breslow (Generalized Wilcoxon)	7.960	2	0.019
Tarone-Ware	8.281	2	0.016

对不同层的湖泊生存分布相等性的检验。

表“总体比较”给出了对三个湖泊生存分布相等性（同质性检验）的检验结果。

分组比较基于以下检验方案。

- H₀: 不同因子等级的生存时间分布是相等的。
- H₁: 不同因子等级的生存时间分布具有相等的差异。

在超出所有三组范围进行的同质性比较（p=0.014 （Chi²=8.536; 对数秩），0.01（Chi²=7.96; Breslow）以及 0.016（Chi²=8.281; Tarone-Ware））中，所有三组检验统计量得出了同样的结果：三组之间的差异在统计上具有显著性。检验统计量的差异与显著性主要是源于发生目标事件的时间点和数量（参见第 4.3 节）。

这一结果强有力地表明，三个湖泊在其生存时间分布方面有所差异。但是，这些检验只能说明三个湖泊在某处存在差异，但无法说明哪些湖泊具体有哪些差异。为了找出这个问题的答案，还需进一步进行两两比较。

4.6.3 利用因子变量与分层变量进行比较（Kaplan–Meier 法）

本节中首先利用与第 4.2.6 节的实例的相同数据展示了分层变量的效应。如果给出的变量“湖泊”不是作为因子，而是作为分层变量，并且不给出其他因子变量，则出现两个方面的变化：

- 生存表输出为分层形式；

■ 无法对分层变量的变量类别进行统计学意义上的比较。

然后，通过一个因子变量与一个分层变量的组合，展示了多种不同的对比方法。

在 SPSS 程序主界面选择以下菜单项：分析→生存函数→Kaplan-Meier。

将变量 TIME 拖动至“时间”栏，将变量 STATUS 拖动至“状态”栏。在工作栏“定义事件”下给出所发生目标事件的编码，例如代码“1”。把变量“湖泊”拖动至“层”栏，单击“继续”按钮。

子窗口“比较因子”。SEE 不能定义为因子（只能定义为其他任何一个离散尺度的变量）。如果“比较因子”一栏下并未给出因子变量，则不能进行生存分布比较。

子窗口“选项”。在“绘图”下选中“生存函数”选项。确保“统计量”一项下选中了“生存表”、“生存时间平均值和中位数”以及“四分位数”选项。单击“继续”按钮。

单击“确定”按钮开始计算。

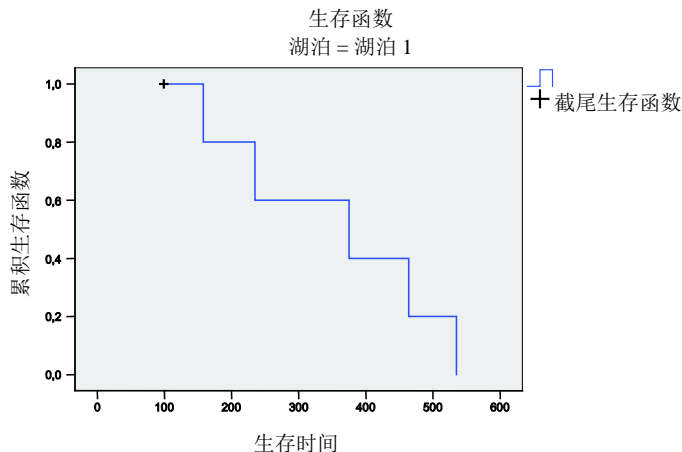
语句：

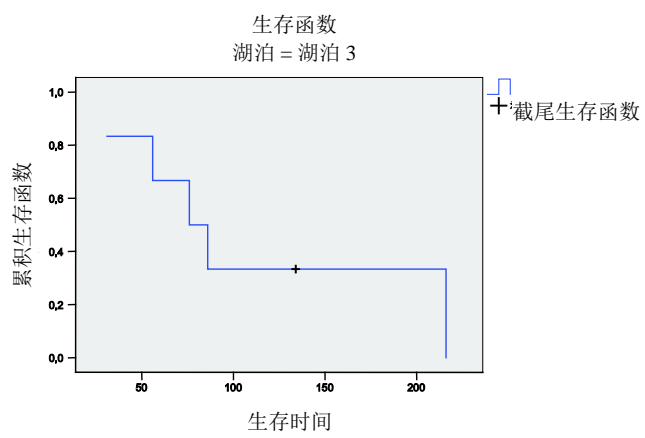
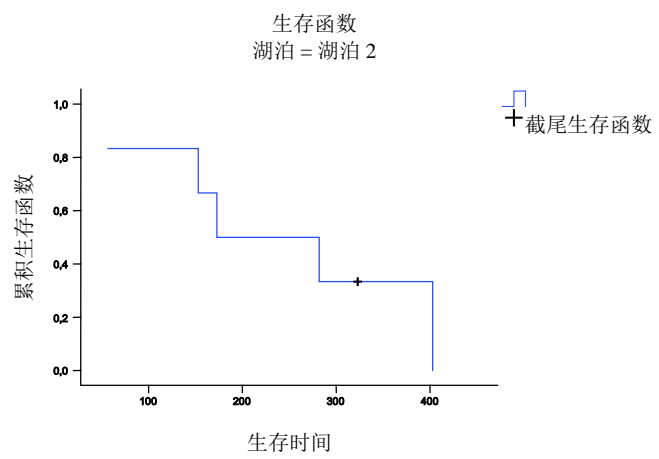
KM

```
TIME
/STRATA=SEE
/STATUS=STATUS(1)
/PRINT TABLE MEAN
/PERCENTILES
/PLOT SURVIVAL .
```

注：KM 命令调出使用 Kaplan-Meier 法的生存分析。TIME 代表生存时间。此外，BY 后缺少一个分组因子。STATUS=后给出的是状态变量（包括目标事件的编码）。PRINT 命令调出“生存表”与“生存时间平均值和中位数”表。PERCENTILES 命令调出“百分位数”表。PLOT 命令删除生存曲线图，包括 SURVIVAL（累积生存函数）、OMS（累积 1 减去生存函数）、HAZARD（累积危险函数）以及 LOGSURV（对数生存函数）。此处不给出 TEST 命令。

输出结果





另一个输出结果与第 4.6.2 节的例子完全吻合（参见“生存表”）。

生存表						
湖泊	时间	状态	某时间点的生存者累积比例		累积事件的 数量	剩余个案的 数量
			估计值	标准误差		
湖泊 1	1 158.000	出现	.800	.179	1	4
	2 235.000	出现	.600	.219	2	3
	3 375.000	出现	.400	.219	3	2
	4 464.000	出现	.200	.179	4	1
	5 535.000	出现	.000	.000	5	0
湖泊 2	1 57.000	出现	.833	.152	1	5
	2 153.000	出现	.667	.192	2	4
	3 173.000	出现	.500	.204	3	3
	4 282.000	出现	.333	.192	4	2
	5 323.000	截尾	.	.	4	1
	6 403.000	出现	.000	.000	5	0

续表

湖泊	时间	状态	某时间点的生存者累积比例		累积事件的 数量	剩余个案的 数量
			估计值	标准误差		
湖泊 3 1	31.000	出现	.833	.152	1	5
2	56.000	出现	.667	.192	2	4
3	76.000	出现	.500	.204	3	3
4	86.000	出现	.333	.192	4	2
5	134.000	截尾	.	.	4	1
6	216.000	出现	.000	.000	5	0

下面的例子说明了，在给定了因子变量和分层变量的情况下，SPSS 有哪四种比较因子等级的方法。上文已经提到，除了每次分别调用的检验方法之外，其他所有输出结果，如“生存表”等都是完全一致的。为了清楚起见，此处输出结果仅限于例子中所要求的对数秩检验。

子窗口“比较因子”：选项“所有层共同”。

```
KM
TIME BY SEE /STRATA=SCHICHT /STATUS=STATUS(1)
/PRINT TABLE MEAN
/TEST LOGRANK
/COMPARE OVERALL POOLED
/PLOT SURVIVAL .
```

总体比较^a

	卡方	自由度	Sig.
对数秩 (Mantel-Cox)	6.485	2	.039

对不同层的湖泊生存分布相等性的检验。

a.针对层经过校正。

对跨越所有层（Strata）的因子组进行比较，最后只输出同质性检验的结果。

子窗口“比较因子”：选项“每个层”。

```
KM
TIME BY SEE /STRATA=SCHICHT /STATUS=STATUS(1)
/PRINT TABLE MEAN
/TEST LOGRANK
/COMPARE OVERALL STRATA
/PLOT SURVIVAL .
```

总体比较

层		卡方	自由度	Sig.
层 A	对数秩 (Mantel-Cox)	5.307	2	.070
层 B	对数秩 (Mantel-Cox)	2.518	2	.284

对不同层的湖泊生存分布相等性的检验。

对每个层（Strata）内部的因子组进行了比较，对每个层都进行了同质性检验。

子窗口“比较因子”：选项“跨层地成对比较”。

```
KM
TIME BY SEE /STRATA=SCHICHT /STATUS=STATUS(1)
/PRINT TABLE MEAN
/TEST LOGRANK
/COMPARE PAIRWISE POOLED
/PLOT SURVIVAL .
```

成对比较 ^a							
湖泊		湖泊 1		湖泊 2		湖泊 3	
		卡方	Sig	卡方	Sig	卡方	Sig
对数秩 (Mantel-Cox)	湖泊 1			1.044	.307	4.927	.026
	湖泊 2	1.044	.307			2.407	.153
	湖泊 3	4.927	.026	2.407	.153		

a. 针对层经过校正。

对所有跨层的因子等级进行比较，最后对每 N 个因子等级输出 $N-1$ 个比较结果。

子窗口“比较因子”：命令“每个层的成对比较”。

```
KM
TIME BY SEE /STRATA=SCHICHT /STATUS=STATUS(1)
/PRINT TABLE MEAN
/TEST LOGRANK
/COMPARE PAIRWISE STRATA
/PLOT SURVIVAL .
```

成对比较							
湖泊		湖泊 1		湖泊 2		湖泊 3	
		卡方	Sig	卡方	Sig	卡方	Sig
对数秩 (Mantel-Cox) 层 A	湖泊 1			.559	.455	2.469	.116
	湖泊 2	.559	.455			2.469	.116
	湖泊 3	2.469	.116	2.469	.116		
层 B	湖泊 1			.486	.485	2.557	.110
	湖泊 2	.486	.485			.485	.486
	湖泊 3	2.557	.110	.485	.486		

此处对每个层（Strata）内部的因子等级进行了成对比较，对每层分别进行了同质性检验。

SPSS 可以通过语句，最多同时进行两个检验。例如：

```
/COMPARE OVERALL STRATA
/COMPARE PAIRWISE STRATA
或者：
/COMPARE OVERALL POOLED
/COMPARE PAIRWISE POOLED .
```

SPSS 过程命令 KM 通过界面选择只能将一个因子纳入分析。例如，除了变量“湖泊”，通过界面选择无法将变量“层”作为第二个因子纳入。

但是，用一个小窍门就可以将变量“湖泊”与“层”同时作为第二个因子纳入生存分析（关于利用 SPSS 的数据管理参见 Schendera 的著作，2005）。

```
do if SCHICHT = 1.
    compute SEE_SCH = SEE .
else.
    compute SEE_SCH = SCHICHT + SEE .
end if.
exe.
list var= SCHICHT SEE  SEE_SCH.
```

这个窍门是：利用 DO-IF 命令把两个分类变量（“SEE（湖泊）”，3 个变量类别；“SCHICHT（层）”，2 个变量类别）合并为一个（SEE_SCH；最多 2×3 个变量类别）。通过这一方法，原则上可以将任意多的变量组合成一个因子，并纳入 Kaplan-Meier 分析。

```
KM
TIME BY SEE_SCH /STATUS=STATUS(1)
/PRINT TABLE MEAN
/TEST LOGRANK
/COMPARE OVERALL STRATA
/PLOT SURVIVAL .
```

4.6.4 Kaplan-Meier 分析的置信区间

用 SPSS 也可以测定 Kaplan-Meier 分析的置信区间，并用图形表现出来。此处介绍的方法中，置信区间的上、下限分别是根据生存函数各个估计值的之和与之差，而这个生存函数是由标准误差与期望的、渐近累积正态分布的置信区间相乘而得的乘积构成的。例如，下面这个例子测算了累积生存概率的 95% 置信区间。以下解释仅针对新出现的步骤。

```
KM
UEBERLEB
/STATUS=UESTATUS(0)
/PRINT TABLE MEAN
/PERCENTILES
/PLOT SURVIVAL HAZARD
/SAVE SURVIVAL SE .
```

注：KM 命令调出 Kaplan-Meier 法的生存分析。在执行 SAVE 命令后，用 SURVIVAL 命令保存累积生存函数，用 SE 命令保存相关的标准误差。SPSS 会自动为新创建的变量分配名称。在本例中，累积生存函数的数值另存为 SUR_1，标准误差的数值则保存为 SE_1。

```
compute UCL = SUR_1 + 1.96*SE_1 .
compute LCL = SUR_1 - 1.96*SE_1 .
if (UCL > 1) UCL = 1 .
exe .
```

```

sort cases by UEBERLEB (A) .
variable labels
    UCL "...的置信区间上限"
    LCL "...的置信区间下限" .
exe.
formats
    UCL (F8.2)
    LCL (F8.2).

SAVE OUTFILE='C:\KM_Konf.sav' .

```

注：置信区间可通过两个 COMPUTE 命令测得。UCL（“upper confidence limit”）命令定义了置信区间的上限，LCL（“lower confidence limit”）命令则定义了置信区间下限。将累积生存函数的各个标准误差值与 1.96 相乘，1.96 就是标准正态分布的 0.025 分位点，定义从这里开始正态分布曲线下的区间面积占总面积的 2.5%。因此，数值-1.96 就确定了占总面积 2.5% 的分位点，+1.96 确定了占总面积 97.5% 的分位点。UCL 与 LCL 加上 SUR_1 值共同定义了累积生存函数周围的 95% 置信区间。如果要使置信区间定义更严格（更窄）或者更宽泛（更宽），则应将 1.96 替换为其他数值。

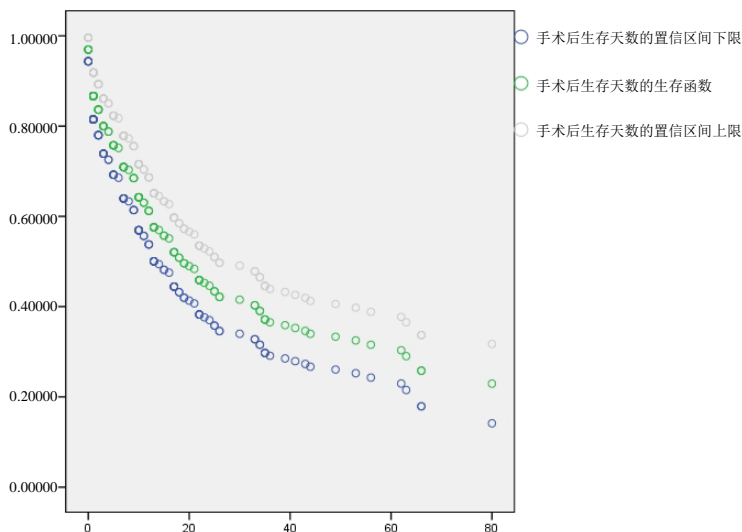
```

GET FILE='C:\KM_Konf.sav' .

GRAPH
/SCATTERPLOT(OVERLAY)=UEBERLEB UEBERLEB UEBERLEB
with LCL SUR_1 UCL (PAIR)
/MISSING=LISTWISE .

```

由于这种方法是基于正态分布的，因此必须具备正态分布这个根本前提条件，也就是样本量至少要达到适当的大小。但也不能认为，这种方法不太适合或者甚至完全不适合较小的样本量。关于推导确定置信区间的其他方法请参见 Hosmer & Lemeshow（1999）的著作。



4.6.5 不带因子的寿命表计算法示例

Kaplan-Meier 法不一定适用于较大的样本量（例如，在第 4.6.1 节中所示）。在某些情况下，寿命表法更适合个案数量较多的情况。

与第 4.6.1 节所述类似，使用寿命表法的下列分析研究 $N=165$ 例患者手术后的生存时间。观察期扩展到超过 100 天，因此，将生存时间划分为多个由 10 天组成的区间，以便进行寿命表分析。

在 SPSS 程序主界面选择以下菜单项：分析→生存函数→寿命表...

将变量 “ueberleb” 拖入 “时间” 栏。在 “显示时间区间” 一项下确定最长观察期（如 100 天）与时间区间（如 10 天）。将变量 UESTATUS 拖入 “状态” 栏。在 “定义事件” 一栏下面输入将要出现的目标事件的编码，例如，代码 “0”。此处无须给定因子（与第 4.6.6 节所述相反）。



子窗口 “选项”。确定已经选定了选项 “寿命表”。在 “绘图” 一项下选定 “生存函数”、“对数生存函数”、“危险函数”、“密度函数” 和 “1 减去生存函数” 选项。子选项 “比较第一个因子的水平”。由于此前没有定义因子，所以这个子选项此时不可用（与第 4.6.6 节所述相反）。单击 “继续” 按钮。

单击 “确定” 按钮开始计算。

语句：

```
SURVIVAL
TABLE=ueberleb
/INTERVAL=THRU 100 BY 10
/STATUS=uestatus(0)
/PRINT=TABLE
/PLOTS ( SURVIVAL HAZARD OMS LOGSURV DENSITY )=ueberleb .
```

注：SURVIVAL 命令调用一个寿命表。在 TABEL=后面，给定了表示生存时间的变量 UEBERLEB。通过 INTERVAL=命令，将从 0 到 100 天的生存时间（见 UEBERLEB）划分为 10 个由 10 天组成的区间（BY 10）。在 STATUS=命令后面给出的是状态变量 UESTATUS。括号内的编码（“0”）代表目标事件。在 PRINT 命令后面，用 TABLE 命令调用寿命表，

NOTABLE 命令会阻止输出寿命表。PLOT 命令可以调用生存曲线图（SURVIVAL：累积生存函数、HAZARD：累积危险函数、OMS：1 减去生存函数、LOGSURV：对数生存函数、DENSITY：密度函数。可以用命令 ALL 调用所有的图）。

输出结果

区间的开 始时间	初始时 的生存 者人数	死亡者 人数	风险承受 者人数	终结事件 的数量	终结 比例	生 存 者 的 比例	生存者在区 间末端的累 积比例	生存者在区 间 终结时的累积 比例标准误差	概率 密度	风险率	风险率的 标准误差
0.000	165	0	165.000	52	.32	0.68	.68	.04	.032	.04	.01
10.000	113	2	112.000	31	.28	0.72	.50	.04	.019	.03	.01
20.000	80	0	80.000	12	.15	0.85	.42	.04	.007	.02	.00
30.000	68	1	67.500	10	.15	0.85	.36	.04	.006	.02	.01
40.000	57	4	55.000	4	.07	0.93	.33	.04	.003	.01	.00
50.000	49	19	39.500	2	.05	0.95	.32	.04	.002	.01	.00
60.000	28	9	23.500	4	.17	0.83	.26	.04	.005	.02	.01
70.000	15	6	12.000	0	.00	1.00	.26	.04	.000	.00	.00
80.000	9	5	6.500	1	.15	0.85	.22	.05	.004	.02	.02
90.000	3	3	1.500	0	.00	1.00	.22	.05	.000	.00	.00

说明：“中位生存期为 19.75”。中位数表明了在某一个具体的时间点，低于和高于中位生存期的个案各占 50%。

这个数值表明，在手术后不到 20 天的时间里，大约 50% 个案的患者死亡。这个结果从内容上来看也是显而易见的，因为手术之后的初期阶段根据以往经验是最危险的一段时间，尤其是对于常规的介入手术而言。在此还要指出的是，使用 Kaplan-Meier 法测定的中位数稍有不同，为 19.00（参见第 4.6.1 节）。

注：如果对于相同的数据，在调查研究结束后只有不到 50% 的个案出现了目标事件，则 SPSS 的图例中有以下说明：“中位生存期为 90.00”。这说明，在最后一个时间区间中至少还有 50% 个案的患者生存。

表“寿命表”包含了许多数据，下面将逐列解释。在测算生存表时使用的个案数 $N=165$ 。

“区间的开始时间”。各个时间区间的开始时间点，也就是如 0、10、...、90。例如，第二个时间区间从第 10 天开始。

“初始时的生存人数”。在开始时已经进入相应时间区间的有效个案的数量。在第二个时间区间中，开始时有 113 个案。这些数值逐次减少。

“死亡人数”。各个时间区间的失效（截尾）个案的数量。由于截尾，所以在第二个时间区间中有两个个案失效。

“风险承受者人数”。患者在手术后有生命危险的个案的数量。这一数值是通过达到相应区间的个案数量减去该区间截尾个案数量的一半得出的。因此，在这个时间区间中的危险个案数量为 $112 = 113 - (2/2)$ 。

“终结事件的数量”。在相应的区间所出现的目标事件的数量。例如，在第二个时间区间出现了 31 个目标事件。

“终结比例”：出现的目标事件的比例。终结事件比例是对于在某个时间区间发生死亡的有条件的估计概率，前提条件是，个案在相应的第 $n-t$ 个时间区间仍然生存。出现的目标事件的比例可以通过危险个案数量与终结个案数量的比例来得出。因此在第二个时间区间中，终结比例为 $0.28 = 31 / 112$ 。

“生存比例”：生存比例通过 1 减去终结比例来得出。因此在第二个时间区间中，生存比例为 $0.72 = 1 - 0.28$ 。PLOTS / (OMS)命令对此输出图像。

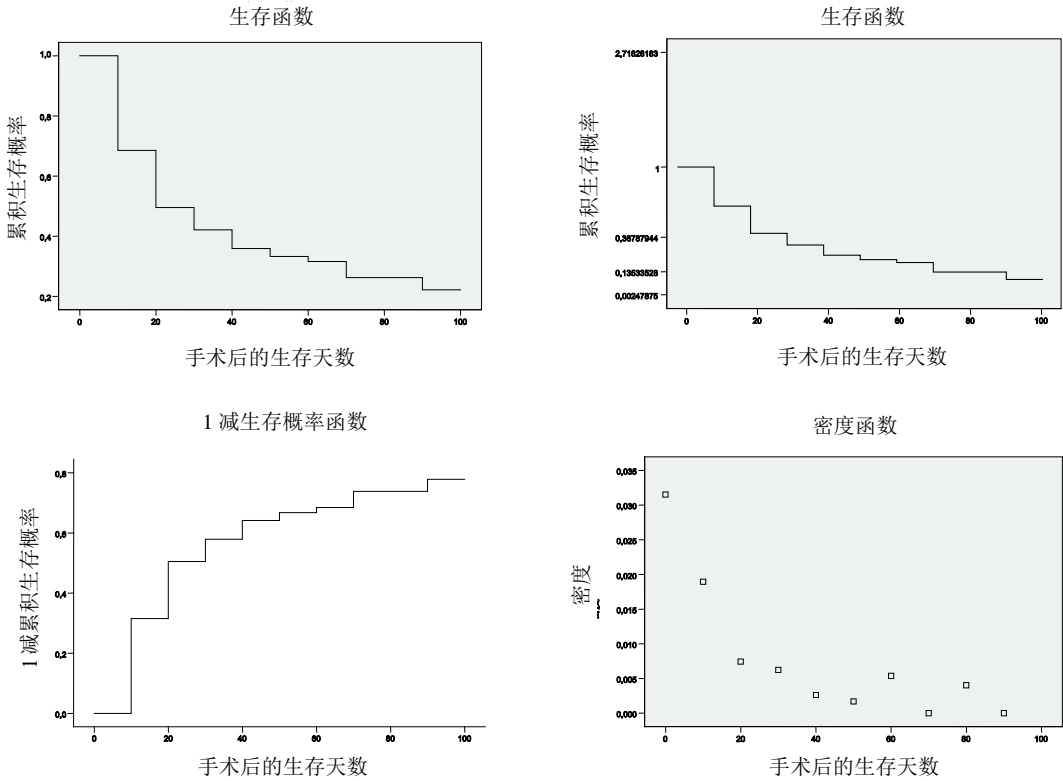
“区间末端的累积生存比例”。生存函数即从表格始端生存到第 $n-t$ 个时间区间末端的估计概率。假设各个时间区间之间的生存概率彼此是不相关的，则在区间末端的累积生存比例可以通过这个时间区间的生存比例乘以前面一个时间区间的生存比例来得出。在第二个时间区间中，区间末端的累积生存比例为 $0.50 = 0.68 \times 0.72$ 。/PLOTS (SURVIVAL) 命令对此输出图像。

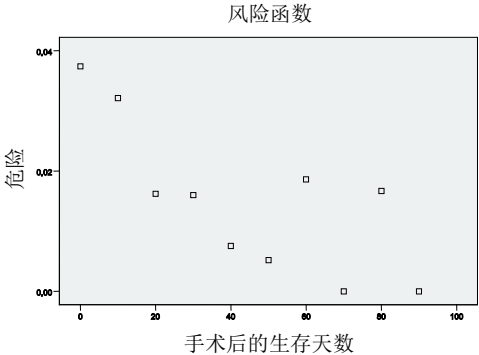
“区间末端累积生存比例的标准误差”。在第二个时间区间中，区间末端累积生存比例的标准误差为 0.04。

“概率密度”。在相应的时间区间发生目标事件的概率估计值。通过/PLOTS (DENSITY) 命令将其可视化。

“风险率”。风险率（又称危险函数/失效率）相当于时间区间中点的估计值，表示只要有一个个案直到这个时间区间仍然生存，就将其作为目标事件出现的概率。在第二个时间区间中，风险率为 0.03。/PLOTS (HAZARD)命令对此输出图像。

“风险率的标准误差”。第二个时间区间的风险率标准误差为 0.01。





注：如果将寿命表的生存函数与 Kaplan-Meier 法（第 4.6.1 节）相比较，则可以确定，Kaplan-Meier 生存曲线与寿命表法的有所不同（因为 Kaplan-Meier 法是基于 $N=165$ 个单独个案，而寿命表分析仅有 10 个时间区间）。此外，Kaplan-Meier 生存曲线还显示截尾，而寿命表曲线则不显示。

实例讨论

累积生存函数。经过了十分艰难的初始阶段（在术后不到 20 天内已经有大约 50% 的个案失效）后，函数曲线稳定成一条缓慢下降的直线，基本上与 X 轴平行。

危险函数。这个图表明，一个目标事件在某个时间区间出现的风险及潜在可能性，前提条件是作为目标事件的个案在这个时间区间仍然生存。因此，前两个时间区间是最关键的。但是风险率也表明，在较晚的时间区间（例如，第 7 个、第 9 个）还是有可能出现失效个案（例如，与联合指征的交互作用，伤口未正确愈合）。在样本容量次优的寿命表中，单个个案可能具有很大的影响。

4.6.6 带有因子的寿命表法计算示例

下面的分析为第 6.4 节的寿命表例子补充了一个因子。尤其是在考虑到一个因子变量的变量类别的情况下研究生存曲线的走向。目的是为了查明两组 $N=165$ 的患者（用药 vs 安慰剂）术后生存曲线的走向是否有所不同。治疗组的患者服用了一种检验样本，对照组的患者服用了一种安慰剂。因此，只有带有分类变量类别的变量才适用于带有一个因子的情况。可以将定量变量分类，从而将其纳入寿命表。在这个例子中，观察期同样超过了 100 天，因此，生存时间在寿命表中被分成了每 10 天一个的时间区间。

在 SPSS 程序主界面选择以下菜单项：分析→生存函数→寿命表...

将变量 ÜEBERLEB 拖入“时间”一栏。在“显示时间区间”下设定观察期（例如 100）的最大值和时间区间的宽度（例如 10）。把变量 UESTATUS 拖入“状态”一栏。在“定义事件”一栏下对出现的目标事件给定编码，例如代码“0”。把变量 MED（其“检验样本”的变量类别为 1，“安慰剂”的变量类别为 2）拖入“因子”一栏。在“定义范围”一栏中定义需要纳入的因子等级的编码最小值和最大值，例如，代码“1”（作为最小值）和“2”（作为最大值）。

子窗口“选项”。确保已选定选项“寿命表”。在“绘图”下方选定选项“生存函数”和

“危险函数”。在子选项“比较第一个因子的层”下单击“全部”，然后单击“继续”按钮。
单击“确定”按钮开始计算。

语句：

```
SURVIVAL
TABLE=ueberleb BY med(1 2)
/INTERVAL=THRU 100 BY 10
/STATUS=uestatus(0)
/PRINT=TABLE
/PLOTS ( SURVIVAL HAZARD )=ueberleb BY med
/COMPARE=ueberleb BY med .
```

备注：SURVIVAL 命令调出一个寿命表。在 TABLE=后面给定生存时间的变量 UEBERLEB。在 BY 后面给定分组因子 MED 和所期望变量类别的范围。通过 INTERVAL=可以将生存时间（参见 UEBERLEB）分成从 0 到 100（0 THRU 100）天、以 10 天为一个区间的 10 个区间。在 STATUS=之后给定状态变量 UESTATUS。括号内是表示目标事件的编码（“0”）。在 PRINT 之后利用 TABLE 调出寿命表。PLOT 输出两个生存曲线图：SURVIVAL（累积生存函数）和 HAZARD（累积危险函数）。与第 4.6.5 节相反，这里的语句后面还有一句 BY MED。通过结尾的 COMPARE 确定比较的方向：“UEBERLEB by MED”根据变量 MED 的变量类别调出对生存时间 UEBERLEB 的比较。

SURVIVAL 语句优于界面选择，尤其是因为它可以让用户进行下列工作。

- 在 TABLES 后面可以为生存时间给定最多 20 个变量。
- 在 INTERVAL 下面设定不相等的时间区间。
- 通过多个/STATUS 命令可以给定多个状态变量。
- COMPARE 可以设定进行比较，这些比较只包含特定的而不是全部的（预设定）因子和控制变量。
- 可以用更适合的近似比较取代精确比较来计算累积数据（例如，通过 /CALCULATE=APPROXIMATE）。

输出结果

寿命表													
一阶控制变量	区间的开始时间	初始时的生存人数	死亡人数	风险承受者人数	终结事件的数量	终结比例	生存比例	区间末端的累积生存比例	区间末端累积生存比例的标准误差	概率密度	概率密度的标准误差	风险率	风险率的标准误差
服药检验样本	0.000	33	0	33.000	4	.12	.88	.88	.06	.012	.006	.01	.01
	10.000	29	1	28.500	4	.14	.86	.76	.08	.012	.006	.02	.01
	20.000	24	0	24.000	1	.04	.96	.72	.08	.003	.003	.00	.00
	30.000	23	1	22.500	4	.18	.82	.60	.09	.013	.006	.02	.01
	40.000	18	2	17.000	1	.06	.94	.56	.09	.004	.003	.01	.01
	50.000	15	6	12.000	0	.00	1.00	.56	.09	.000	.000	.00	.00

续表

一阶控制变量	区间的开始时间	初始时的生存人数	死亡人数	风险承受者人数	终止事件的数量	终止比例	生存比例	区间末端的累积生存比例	区间末端累积生存比例的标准误差	概率密度	概率密度的标准误差	风险率	风险率的标准误差
	60.000	9	3	7.500	0	.00	1.00	.56	.09	.000	.000	.00	.00
	70.000	6	2	5.000	0	.00	1.00	.56	.09	.000	.000	.00	.00
	80.000	4	2	3.000	0	.00	1.00	.56	.09	.000	.000	.00	.00
	90.000	2	2	1.000	0	.00	1.00	.56	.09	.000	.000	.00	.00
安慰剂	0.000	132	0	132.000	48	.36	.64	.64	.04	.036	.004	.04	.01
	10.000	84	1	83.500	27	.32	.68	.43	.04	.021	.004	.04	.01
	20.000	56	0	56.000	11	.20	.80	.35	.04	.008	.002	.02	.01
	30.000	45	0	45.000	6	.13	.87	.30	.04	.005	.002	.01	.01
	40.000	39	2	38.000	3	.08	.92	.28	.04	.002	.001	.01	.00
	50.000	34	13	27.500	2	.07	.93	.26	.04	.002	.001	.01	.01
	60.000	19	6	16.000	4	.25	.75	.19	.04	.006	.003	.03	.01
	70.000	9	4	7.000	0	.00	1.00	.19	.04	.000	.000	.00	.00
	80.000	5	3	3.500	1	.29	.71	.14	.05	.005	.005	.03	.03
	90.000	1	1	0.500	0	.00	1.00	.14	.05	.000	.000	.00	.00

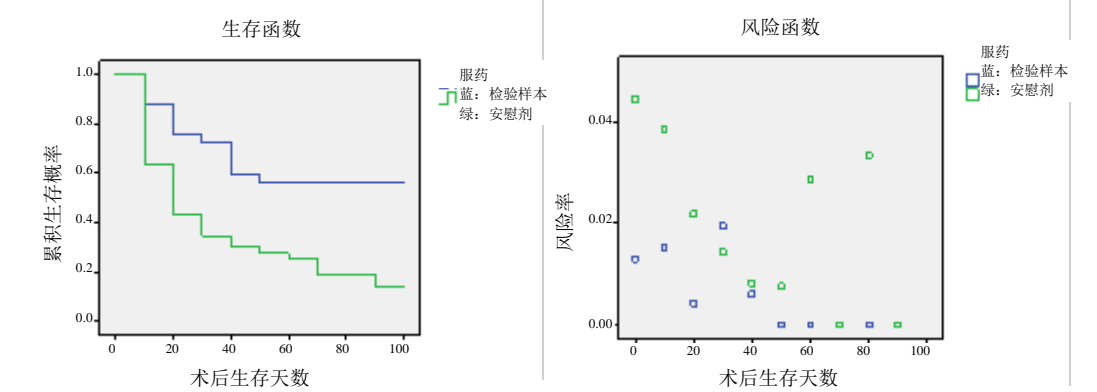
这些寿命表在结构与第 4.6.5 节中的例子一致。唯一的不同之处是，这个表格是按照给定因子的变量类别划分的。对于因子的每个变量类别（“检验样本”，“安慰剂”）都输出了一个子表。

中位生存期

一阶控制变量		中位生存期
用药	检验样本	90.00
	安慰剂	16.63

如果给定一个因子，就会单独为“中位生存期”输出一个表格。该表表明了每个所给定因子的变量类别的中位生存期。中位数表明了一个具体的时间点，每个因子等级的个案分布在这个时间点以上和以下，分别有 50% 的个案具有超过和低于中位生存期的生存时间。

新的样本看起来非常成功。90 天后还有超过 50% 的患者仍然生存，而安慰剂一组患者的生存时间明显要糟糕得多。从第 6 个时间区间开始，检验样本一组的生存曲线保持恒定不变。因此，引入一个因子就能从另一个角度观察药物的疗效（参照下文）。



由于一种效应（在这里：死亡）是定义为负向的，所以在累积生存函数的图中可以清晰地看出，检验样本一组的生存曲线一直高出安慰剂一组，并且呈缓慢下降的趋势。

检验样本呈现出明显更有利的效应，也就是说，生存概率超过安慰剂一组。从第 6 个时间区间开始就不再出现死亡个案（左图），而在对照组中，死亡危险甚至增大了（右图）。下面的表则表明，这两个组之间的这个差异是否达到了统计显著性。

控制变量的比较值：med

总体比较 ^a		
Wilcoxon-(Gehan)统计量	自由度	Sig
11.165	1	0.001

a. 比较是精确的。

“总体比较”表包含了对所有小组的生存时间进行一致性检验的结果。数值为 11.165（df=1，p=0.001）的 Wilcoxon（比分）统计量表明，“检验样本”和“安慰剂”两个组在统计数据上有显著差异（因为 $p < 0.05$ ）。

如果分为超过两个组，则一个显著的 Wilcoxon（比分）统计量只能说明，两个组在统计数据上有显著差异，但是检验并不能表明存在差异的是哪两个组。这些信息只能通过成对的检验才能得出。

新的检验样本对术后生存时间明显有积极的影响。

结论

寿命表分析与 Kaplan-Meier 分析不同，它适用于数据量很大的情况。

在界面选择的两种方法中，仅限于对最多两个分类变量（Kaplan-Meier：因子变量和分层变量、寿命表：两个因子）建立模型。只有通过编程才能使建立的模型实现复杂的用途，例如，将定量变量分类、将多个分类变量合并为一个或者把所选择的因子纳入比较。

但是如果要将定量变量，尤其是时间相依变量纳入分析模型，则应选择 Cox 回归法（参见第 4.7 节）。

补充说明：寿命表中类别宽度的意义

采用寿命表法时，原则上可以将给定的生存时间完全自由地划分为各个区间。

对观察结果进行分类（分级、分组）时既可以遵循纯粹的形式标准（例如，所包含元素的数量相同、时间区间宽度相同，分布相同，区间的数量相同），也可以只遵循内容标准（例如，根据特定的质量特征，如临床检验时设定的极限值）。然而，相同宽度的时间区间并不一定意味着在这些时间区间内的元素数量也相同，同时也不意味着这些时间区间的内容就一定具有可比性。在实践中要同时运用形式标准和内容标准，在有疑问的情况下，优先使用内容标准。

由于存在隐性的加权，形式或语义上不同的时间区间宽度可能会导致信息的扭曲。如果只遵循形式标准，最晚在对其进行内容上的解释时就会暴露其缺点。因此，不能通过一个公式，即完全从理论角度对时间区间进行定义，而是要采用理论推导并同时检验的方法进行定义。对

于按照形式标准，也就是通过理论推导来定义的时间区间来说，同样适用风险评估的运作方式：如果时间区间太宽，则可能有重要的分布特征被遮盖的风险；而如果时间区间太窄或太小，则对这些时间区间就对随机效应或离群值太敏感。

在 SPSS 中，可以有针对性地通过条件（IF）、COMPUTE 或 RECODE 来定义时间区间（参见 Schendera, 2005）。

补充说明结束

4.6.7 计算生存分析的首要条件

生存分析不需要普通的前提条件，例如，多变量正态分布、线性或同方差性。

1. 零点的选择是分析时间相关数据的基础所在。通常，建议将一种效应影响的始点（例如，大学生活或者一项治疗开始的日期）作为零点。某个疗法随机分配的、相同的时间点（例如，一项调查研究的开始日期），或者对风险率（风险）影响最大的时间点（如某个诊断的时间点）更适合作为零点。
2. 在调查研究期间，出现了时间相依的事件。即肯定至少有一个时间变量，它可以描述一个事件的发生或者截尾的时间点。
3. 生存分析的时间值必须是比率尺度的（零点！）正数值，并且都采用相同的时间单位（天、年、月、小时等）。负数值将被删除。
4. 时间变量必须在其尺度标定上是连续的，因子变量和分层变量必须是定类的。状态变量可以是定类的或连续的。
5. 如果要考虑到截尾，就需要有一个变量来区分截尾和非截尾。通常情况下，将截尾编码为“0”，将没有截尾的数据（也就是事件）编码为“1”。然而，根据作者或数据分析师不同，还可以采用另外一种编码方式，因此必须对其进行检验。
6. 一个元素只能有一个失效值，一个元素只能有一个截尾。同一个元素不可能从初始数据中失效两次，同一个元素不能两次被截尾。
7. 目标事件的概率只允许与因果因子（例如治疗效果）和时间因子有关。在这里认为，因果因子和环境因子在研究过程中是不变的，并对生存函数产生恒定的影响。如果因果因子或者环境条件不断变化，则违反了生存分析这个隐含的基本假设。如果各个因素是在不同的时间被纳入初始数据（例如，患者的治疗起始时间不同），那么应通过敏感性分析确定在这些元素之间不存在差异。
8. 隐性因子（分层变量）。目标事件的概率只允许取决于因果因子（例如，治疗效果）和时间因子。如果这个概率还有其他因子，如不同的治疗开始时间，则有可能会影响生存时间以及截尾。这种所谓的“隐性因子”损害了数据的独立性，即（观察的、截尾的/未截尾的）样本数值不再是来自分布状况相同的一个总体。与建立组块类似，可以通过建立分层变量检验隐性因子，例如，将所有患者划分为治疗开始时间类似的各组。
9. 截尾。随机的截尾肯定是毫无信息价值的（即所谓的“随机截尾”）。截尾不得与分层变量、因子变量或时间变量的各个变量类别有任何关系。例如，截尾不应与生存时

间有关，不同组中的截尾模型应相似。随机截尾的比例应尽可能小，并且必要时通过适当的协变量可以对其进行检验。如果有大量的截尾，则表明这项调查研究出现了问题（例如，结束太早或截尾模型不当等）。在截尾的和未截尾的事件之间，不应发现有任何系统性的差异，否则这种偏误会影响调查研究的结果。各组的截尾比例应大致相同。如果各组的截尾比例不同，则可能表明有偏误，从而破坏比较的结果。例如，在进行两种疗法的比较时，由于研究的中断，接受次选疗法的患者比接受首选疗法的患者更快地在这项调查研究中失效。

10. 对曲线走向和检验的解释。如果需要比较的各组的曲线走向是相互不平行或者不相交的，那么就不适合采用对数秩检验、Breslow 检验和 Tarone-Ware 检验作为检验方法。在解释检验结果时，要考虑到样本量、事件和截尾的数量和分布，以及所选择检验方法的加权重点。例如，曲线走向应在截尾方面好、不显著，在相互比较的各组中，曲线走向应是近似的。
11. 样本量。对数秩检验、Breslow 检验和 Tarone-Ware 检验是基于卡方统计量的，因此，如果样本量越大，检验就越精确。模型中纳入的协变量越多，或者说模型的性能越强，则就需要越多的个案。在对两个或多个组进行比较时，最好各个组的样本量都相等，至少也要近似相等，否则零假设就可能会被错误地拒绝（参见 Klein & Moeschberger 著作，2003，214）。每个组的测量值数量影响各个生存函数所做的估计。如果想对估计精确度不同的生存函数或生存曲线进行比较，那么没有多少意义。如果样本量太小或者各组的样本量相差太大，那么就不再适合使用基于卡方法的渐近法了。
 例如，对于寿命表法，建议最小样本量为每个时间区间 $N=30$ 。对于 Cox 回归适用的规则是，创建模型时每个因子或者协变量要考虑到 5 到 10 个目标事件（参见 Klein & Moeschberger 著作，2003）。例如，Eliason（1993）建议，对于有 5 个协变量的模型，应将最小样本量设为 $N=60$ 。没有截尾的数据失效得越少，所需的样本量可能就越大。
12. 发生目标事件。在进行生存分析时，重要的是检验这两点：所感兴趣的目标事件是否（以足够的频率）发生？目标事件的频率究竟是符合总体，还是不成比例？
13. 因子的数量。Kaplan-Meier 分析中，每次只能给定一个因子变量和分层变量；在使用寿命表法时，只能给定两个因子。如果要将多个因子输入模型，那么必须事先通过数据管理将这些因子合并成一个因子（参见第 4.6.3 节）。
14. 缺失值。缺失值在建立一个生存时间模型的过程中可能会导致问题。生存时间模型的理想条件是没有任何数据缺失。如果数据是完全随机缺失的，则具体的缺失程度决定了分析时还留有多少百分比的数据，这仍有可能导致出现问题。如果通过合理的考虑，发现缺失值以某种方式与目标变量相关，那么只要从模型中剔除这些缺失值，模型的解释和建立就会出现。例如，通过（a）建模的方式，即通过一个指示缺失值的指标和（b）重建的方式，对缺失数值进行分析（Missing Value Analysis），就可以将缺失数据重新引入模型，但是只能在这个前提条件下：这些缺失数据的编码、重建和模型集成是合理的并且可追溯的（参见 Schendera，2007）。如果缺失值集中在一个变量上，则或许也可以从分析中剔除这个变量。

如果多个分类变量的组合会导致产生很多空白单元格，则既可以从模型中剔除不重要的分类变量，也可以将分类变量的各个变量类别合并起来。为了保证数据完整性，应小心地将各个变量类别合并起来。

15. 因子等级的线性趋势（只是 Kaplan-Meier）。“因子等级的线性趋势”检验仅适用于因子等级呈现自然顺序的情况（例如，当因子等级是由不同的、在理想情况下等距大小的剂量组成的）。

4.7 Cox 回归

4.7.1 Cox 模型简介和背景知识

Cox 模型（又称 Cox 回归、比例风险模型）是一种最常用的生存数据分析方法，与 Kaplan-Meier 法和寿命表法相反。例如，要研究多个定量影响变量（协变量）对（截尾的）生存时间的效应时，就可以使用 Cox 模型。可建模的因素就是不同组的配置，比如在临床试验中的治疗组和对照组。这样，在同时考虑到其他相关协变量的情况下，Cox 回归就可以作为一种真正多变量的生存分析方法，从患者的生存率角度来研究疗效的相关性和有效程度。再根据预测的有关变量进行调整，通常即可实现更精确的估计（参见 Klein&Moeschberger, 2003, 250-253; Allison, 2001, 14）。Cox 回归的标准模型研究的是由一次性出现的事件造成的风险。

Cox 回归的第一种用途是研究由几个一次性出现的等价（竞争）事件造成的风险。也就是说，是否以及什么时候会因为两个（或者多个）原因导致目标事件（如死亡）。由于这两个或者多个先前确定的原因，只允许发生一个事件（“非此即彼”原则）。这种竞争风险生存分析（competing risks survival analysis，又称多目的模型）的对象有，例如，由于肿瘤或者感染导致死亡、由于更换供应商或者收费表导致取消合同，以及由于违反假释规定或者重新犯罪而将假释人员重新羁押（参见 Kleinbaum & Klein, 2005, 第 9 章）。Cox 回归的第二种用途是研究重复出现的事件，例如，酒瘾患者中断治疗、重新羁押假释人员、签订合同后客户在一些售后服务网点投诉，或者在治疗过程中患者的心脏病反复发作。这种方法被称为复发事件生存分析（recurrent event survival analysis，又称为重复事件模型、多周期模型）（参见 Kleinbaum & Klein, 2005, 第 8 章）。SPSS 目前无法对这种模型进行分析。因此，本章没有阐述竞争风险生存分析和复发事件生存分析。

Cox 模型是很多专业人员分析生存数据的首选工具（参见 Kleinbaum & Klein, 2005, Altman, 1992; Bland, 1995; Collett, 2003²; Guggenmoos-Holzmam&Wernecke, 1996; Kahn&Sempos, 1989）：

- 可以研究多个定量影响变量（协变量）对（截尾的）生存时间的效应。
- Cox 模型不以生存时间特定的分布形状为前提条件，其基本的因变量为风险率。
- 为了得出可靠而稳健的结果，Cox 模型只需极少数量的假设就足够了（如风险率、比率、函数）。
- 只要满足了前提条件，Cox 模型就比生存分析的参数分析法更加稳健。

“危险”的概念

危险函数是 Cox 回归的核心概念之一，下面对其详细阐述。Cox 回归在主要几个方面与传统的调查研究设计方案都有不同。例如，在一个队列研究设计方案（如前后对比）中，观察时间是设定为固定的，即对于所有元素都是同样长的；由此可以直接推导出预测的时间范围。在进行生存分析时，通常只有观察时间的起点是固定的，而终点是不固定的。一般不给出同样长的观察时间，而是各个元素的观察时间各自不同。现在，利用风险或者危险函数的概念，即可将不同长度的观察时间转化为一个共同的函数。根据这个函数可以推导出某个“普通”元素的目标事件（如购买、痊愈、解除合同、死亡）在某个时间发生的风险，前提是这个目标事件在此之前还没有发生。风险值越大，则发生这个事件的风险就越高。风险率是衡量一个组内部的风险的度量；风险比是第一个组的风险率与第二个组的风险率的比例。风险具有比例性的基本假设还包括，在一个风险比中，两个组的风险率可以解释为相互成一定的比例，并且随着时间的推移保持不变。

根据模型中的其他协变量进行调节，Cox 模型可以估计出影响生存时间的效应（如药物、治疗）。这样，Cox 模型就可以针对所感兴趣的目标事件测定出某个元素（个案、人等）的风险率（又称“危险”、“潜力”，风险不是概率）。其前提条件是，这个元素的所有协变量都已经同时给出了数值（Kleinbaum & Klein, 2005, 94-103）。

作为估计过程依据的模型是由 $h(t)=h_0(t)\times\exp(\beta_1X_1+\beta_2X_2+\dots+\beta_iX_i)$ （参见 Kleinbaum & Klein, 2005, 94-96）定义的。 $h(t)$ 是危险函数。 $h_0(t)$ 是基线危险函数，只相依赖于时间，不得为负值。因此，基线风险 $h_0(t)$ 是一个非特定的函数。由于 $h_0(t)$ 的这个特性，Cox 回归也被称为半参数模型。不受协变量的影响， $h_0(t)$ 估计出发生目标事件的风险，单位为 t 。 \exp 是指数函数。 $\beta_1x_1+\dots+\beta_kx_k$ 是时间不相依协变量 X_1,\dots,X_i 的线性函数，将这个函数变为指数函数，以避免出现负值而无法处理。 X 表示 i 个协变量的集束（“向量”），也就是对风险率可能产生影响的因素。 β_1,\dots,β_i 表示将要估计的协变量回归系数。

现在，Cox 模型确定了这一点：风险率在某个时间点 t 时，仅仅是两个具有相乘关系的变量的乘积，也就是生存时间和协变量向量的乘积，水平为 $h_0(t)\times\exp(\beta_1X_1+\beta_2X_2+\dots+\beta_iX_i)$ 。其中， $h_0(t)$ 代表生存时间， $\exp(\beta_1X_1+\beta_2X_2+\dots+\beta_iX_i)$ 代表协变量的向量。第一个变量 $h_0(t)$ 只涉及“影响因素”时间，不涉及协变量，因此被称为基线危险函数。第二个变量是 X_i 的线性和的对数函数（ \exp ，见下文），通过给出的所有协变量 X 计算出这个线性和；对数函数只涉及协变量，不涉及“影响因素”时间。现在，对于 Cox 模型重要的是这个假设：基线危险函数不相依赖于时间，但并不是给出的协变量 X 。相反，对数函数只是基于协变量 X ，与时间不相关。因此，这些协变量被定义为时间不相依协变量（Cox 模型 1）。通常，时间不相依协变量包括了生理性别等变量，或者其他的分层变量。必须不断地对定距协变量是否不相依赖于时间进行检验。生存时间和协变量向量之间相乘就得出了危险函数，从中可以推导出各自的风险，即只要一个普通元素的目标事件还没有发生，则这个目标事件在某个时间点的发生概率有多大。此时，风险值等于基线风险与协变量效应的乘积。

风险具有比例性的假设

在风险具有比例性这个假设下，由于可以分解为基线危险函数和对数函数，所以可以用回归分析方法将这些协变量参数化。回归系数可以作为关联强度的度量估计出来，并且与“普通

的”回归类似，被解释为衡量各个协变量意义的度量。如果一个协变量的数值发生变化，则系数 β 反映了当影响变量变化了一个单位时风险的预期变化。

现在，如果所有时间不相依协变量都等于零，则 Cox 公式可以简化为基线危险函数 $h_0(t)$ 。Cox 公式的对数部分成为零的对数，即等于 1。换言之，Cox 回归首先假设模型中没有协变量，计算出第一个初始变量，即基线风险 $h_0(t)$ 。Cox 模型唯一的前提条件是，不同变量对于生存的效应在一定时间内保持不变，并且叠加（参见第 4.7.8 节）。如果具备了这个前提条件，则 Cox 模型可以查明解释性变量，也就是协变量对于生存时间的效应。如果基线风险只相依赖于生存时间，则所有时间点的协变量效应是一样的。因此，两个任意个案在任意时间点的风险比就等于其协变量的效应的比例。这样，Cox 模型就实现了这个目标：在模型假设最少的前提下，可以同时估计出多个协变量对生存时间的影响。

Cox 模型的基本假设，即两个组（例如治疗组 T 和对照组 K）的危险函数的商（也就是两种风险的商）随着时间的推移保持不变，可以不相依赖于时间地推导出（经过调节的）风险比（例如 Kleinbaum & Klein, 2005, 108, 215）：风险比 $= h_T(t) / h_K(t)$ 保持不变。在风险具有比例性这个假设下，可以推导出两个结果：a) 风险的比例性在一定时间内保持不变。b) 一个个案的风险始终与任何一个其他个案的风险成比例。从图形上看，这个假设应表现为随着时间的推移平行于对数危险函数曲线（对此也可参见用于检验前提条件的二次对数生存率图）。

风险具有比例性这个基本假设既是 Cox 模型的优点，同时也是缺点。优点是，（只要风险具有比例性的假设成立）从 Cox 模型可以推导出：一种（协变量）效应无论是通过与参考组的比较，还是随着时间的推移都可以解释为保持不变。相应的，治疗组 T 始终成比例地好于对照组 K，就如同一个个案的风险始终与任何一个其他个案成比例。但是由于同样的原因，必须批判性地看待风险具有比例性的假设。因为各个函数完全平行的这个假设对于待调查的研究对象来说不一定是现实的（参见比例性假设检验一节中的论述和示例）。相反，根据研究对象不同，结果可能是各个函数相互交叉或者至少相互靠近。在这种情况下，风险具有比例性的假设不成立。因此，在检验 Cox 模型的模型假设时，风险具有比例性的假设具有非常重要的意义，之后有单独一节对此进行阐述。因此，相对风险保持不变的假设通常只适用于特定的或者有限的观察周期。超过这个观察周期的结果不得用 Cox 回归分析的方法予以解释。在风险不具有比例性的情况下，不得使用 Cox 模型；必要时可以使用其他替代方法。

与寿命表法、Kaplan-Meier 法和回归法的比较

在使用 Cox 回归法时，可以很方便地将一个定量协变量对于生存时间的影响纳入一个模型中；与其相比，在使用寿命表法或 Kaplan-Meier 法时只能实现近似地、并且有信息丢失的分类。Cox 模型是一种真正多变量的回归分析方法，也就是说，在寿命表法或 Kaplan-Meier 法无能为力时，用 Cox 模型也可以实现建模。Cox 建模也意味着，协变量（包括生存时间的，例如在使用寿命表法时）的定距尺度的信息不会由于分类而损失。此外，也可以同时分析一个或者多个定量变量和定性变量的影响，以及可能产生的交互作用的影响。因此，Cox 模型可以参数需求小并且有效地分辨生存时间的行列式。此外，在考虑到截尾个案的情况下，可以确定定量变量和/或定性变量（协变量）对生存时间的影响，并通过回归分析组件将其预测出来。

与多重或者逻辑回归模型类似，Cox 模型的目的是同时估计出不同协变量的影响。截尾数据使得在这里运用传统的方法（如回归分析）变得困难甚至完全不可行。因此，回归分析方法

不太适用于分析生存时间。由于回归分析方法无法区分目标事件和截尾，因此如果运用这些方法，就会丢失截尾数据这样的额外信息。

在下列情况下，无法运用传统的回归模型。

- 没有考虑到对观察数据的截尾。
- 不是为了将时间相依预测变量纳入模型而设计的（参见自相关的问题）。
- 人们感兴趣的因变量（也就是生存时间）通常不具有适当的分布形状，尤其是相同数值的局部集中会扭曲参数估计值。

总而言之，在下列情况下建议使用 Cox 模型。

- 要确定协变量对生存时间的影响。
- 没有关于风险随着时间的演变过程的信息。利用 Cox 模型对风险的分布进行预测是一种半参数的解决方案，因为生存时间的分布无须是已知的。因此，对基线危险函数的预测是根据数据推导出来的。相反，在使用参数类方法时，对函数的选择和预测是根据理论推导出来的。
- 用参数类方法不能适当地对已知的风险分布进行建模。当生存时间具有已知的分布，例如 Weibull、指数、对数成长或者伽马分布时，可以使用参数模型。这里不再阐述参数模型（参见 Kleinbaum & Klein, 2005, 第 7 章；Box-Steffensmeier&Jones, 2004, 66）。
- 在检查随着时间的变化时，只对协变量的大小和作用方向感兴趣。

Cox 模型 1 设定的前提条件是时间不相依协变量。扩展的 Cox 模型也可以考虑到与时间相依的协变量。这个模型 2（带有时间相依协变量的、扩展的 Cox 模型；对此参见 Therneau&Grambsch, 2002²）在建模时考虑到了与时间共变的变量。时间相依协变量，如包括年龄、年收入、记忆力等变量，如果在建模时没有适当地考虑到这些协变量的效应，则分析结果就会产生严重错误。在 Cox 回归过程中，将要执行的、针对时间不相依性或者风险比例性假设的检验（参见第 4.7.6 节）决定了可以使用哪个模型。

补充说明：基本估计方法的背景知识

Cox 回归的估计方法（参见 Kleinbaum & Klein, 2005, 98-100; 111-115）源自于偏似然法。偏似然法之所以得名，是因为它只使用一部分数据，即所发生目标事件的观察序列，只针对达到目标事件的个案测定风险，不考虑截尾。在使用偏似然函数时，在 Cox 模型中首先需要确定基线危险函数的似然值（也就是没有变量的影响）。在下一步，测定在任意一个时间点的目标事件的风险。然后测定似然比，即模型的-2 对数似然值。通过将未指定的函数分段最大化，利用一种迭代的解决方案（牛顿迭代法）测算出所需的估计值，在样本量足够大时呈现相合渐近正态分布，并且有效性的损失很小，也就是说，基本上相当于一种十分稳健的、利用参数的解决方案。如果测算出了似然估计值，则可以将推论统计学方法应用于所测算的风险比（例如利用似然比或者 Wald 检验）。

Cox 回归的计算和解释

建模的出发点与寿命表法或 Kaplan-Meier 法类似，即一个生存时间的变量和一个定义目标事件的状态变量必须是已知的。在 Cox 回归中，必须至少给出一个协变量。两个或者更多协

变量可以是任意的尺度水平，在这里除了其各自的影响之外，还对其交互作用进行研究。此外，还可以设定一些同样可以差异化地观察模型的变量块。如果各个变量可能具有类似的作用方向，如生活习惯、疾病等，则使用变量块更为有利。为了清晰起见，以下关于 Cox 回归的所有示例都使用了“进入”法（ENTER）。这个同样逐步推进的方法的原理与本书第 2 章和第 3 章中所介绍的方法一样。

本章中的 SPSS 示例涉及下列分析模型：

- 4.7.2 带有定量协变量的 Cox 回归
- 4.7.3 带有二元协变量的 Cox 回归（ $k=2$ ）
- 4.7.4 带有分类协变量的 Cox 回归（ $k>2$ ）
- 4.7.5 针对交互作用的 Cox 回归
- 4.7.6 检验 Cox 回归的前提条件
- 4.7.7 带有时间相依的定量协变量的 Cox 回归
- 4.7.8 Cox 回归的特定前提条件

4.7.2 带有定量协变量的 Cox 回归

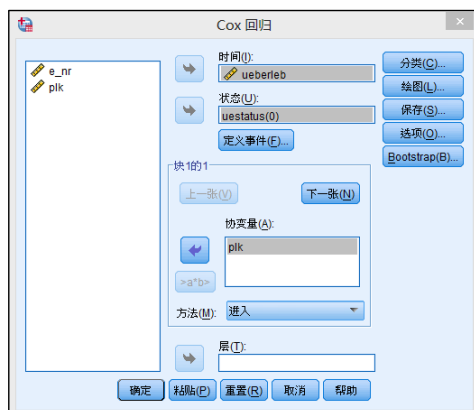
提出的问题

应在考虑到一个定量协变量 PLK 的情况下，研究手术后生存时间的变化过程。目的是找出抗体（PLK）的数量是否会对 $N=165$ 个患者的术后生存时间产生影响，以及如何产生影响。最后进行一个简单化的前提条件检验，即检验风险是否具有比例性，或者抗体的数量是否与生存时间不相依。

在 SPSS 程序主界面选择以下菜单项：分析 → 生存函数 → Cox 回归...

将变量“UEEBERLEB（生存）”拖入“时间”栏。将变量“UESTATUS（状态）”拖入“状态”栏。在“定义事件”一栏下给出将要发生的目标事件的编码，例如“0”。将变量“PLK（抗体）”拖入“协变量”栏。在“方法”一项下选定“进入”。

在选定了“进入”法的情况下，用一个步骤一次性纳入所有变量，在一个步骤之后，这个方法停止运行。只有当影响变量的数量和判别效率是已知的或者至少是固定设置的情况下，才能使用“进入”法。如果判别潜力不明或者要测定一个所含有变量尽量少的有效预测模型，则应使用逐步法。



子窗口“分类...”：不进行任何设置。

子窗口“绘图”：选定选项“生存函数”和“危险函数”，不进行下一步的设置。

子窗口“保存...”：勾选“保存偏残差”。

利用 DfBetas 可以进行离群值分析。利用一个 ID 变量在散点图上截取 DfBetas（参见 Schendera, 2007；不再阐述），不进行下一步的设置。在执行了 Cox 回归后，在 X 轴上调出生存时间的散点图，在 Y 轴上调出 PLK 的偏残差（不再阐述）。

子窗口“选项...”：在“模型统计量”一项下选定“Exp(B)的置信区间”，应用预设设置的数值 95。对于进入模型和剔除模型的预设设置数值，在“逐步法概率”一项下不要做任何改动。在“显示模型信息”一项下选定预设设置的“在每个步骤”。在“迭代次数”一项下，用预设设置的数值 20，即这个模型最多只能迭代 20 次。选定选项“显示基线函数”（对于时间相依协变量不可用）。单击“继续”按钮。

单击“确定”按钮开始计算。

语句：

```
COXREG
  ueberleb /STATUS=uestatus(0)
  /METHOD=ENTER plk
  /PLOT SURVIVAL HAZARD
  /SAVE PRESID
  /PRINT=CI(95) BASELINE
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) .

GRAPH
  /SCATTERPLOT(BIVAR)=ueberleb WITH PR1_1
  /MISSING=LISTWISE.
```

备注：COXREG 命令调出一次 Cox 回归，然后立即给出生存时间变量（UEBERLEB）。在 STATUS=后面给出状态变量 UESTATUS。括号内的代码（“0”）代表目标事件。在 /METHOD 一项下，利用 ENTER 命令调出“进入”法。进入法在第一个步骤后立即停止运行。在 ENTER 后给出所需的效应，在本例中就是定距变量 PLK。PLOT 调出生存曲线的两个图：SURVIVAL（累积生存函数）和 HAZARD（累积危险函数）。在 SAVE 一项下，在 PRESID 后面将模型的偏残差保存在活动数据集中。在 PRINT 后面，利用 CI（95）调出比值的置信区间，利用 BASELINE 调出表格“基线函数”。

在 CRITERIA 后面，可以将迭代计算 Cox 回归模型的所需参数传递给 SPSS。需要给出的变量主要取决于在 METHOD= 一项下指定了哪种方法。只要达到目标标准（例如 ITERATE、BCON 或 LCON），迭代就结束。利用 PIN 和 POUT 确定，根据哪些参数决定变量进入模型还是从模型中剔除。通过 PIN（0.05），设定了决定一个变量是否可以进入模型的数值。如果一个变量的计分统计量的概率小于进入值，则这个变量进入模型；给出的进入值（PIN）越大，变量就越容易进入模型。利用 SPSS 预设设置的 0.05 被认为是相对保守的；为了使潜在的有关影响变量进入，最多可以接受 0.20。利用 POUT（预设置 0.10）指定一个数值，根据这个数值衡

量是否从模型中剔除一个变量。如果概率超过剔除值，则根据有条件的 LR 或者 Wald 统计量将这个变量剔除；给出的进入值（POUT）越大，变量就越容易留在模型中。进入值必须小于剔除值。ITERATE（20）通过一个为正数的整数确定的迭代的最大次数，例如在这里是 20。通过迭代法估计出似然比系数（-2 对数似然值）。如果达到迭代的最大次数，则在达到收敛之前停止迭代。

输出结果
Cox 回归

个案处理评估			
		N	百分比
分析时的可用个案	事件 ^a	116	70.3%
未使用的个案	已截尾总共缺失了数值的个案	49	29.7%
	负数时间的个案	0	0.0%
	一个班次中出现最早事件之前的截尾个案	0	0.0%
	总数	0	0.0%
		165	100.0%

a. 因变量：自手术日期之后的生存天数。

表“个案处理评估”的“分析时的可用个案”一栏给出了目标事件的数量（“事件”， $n=116$ ，70.3%）、截尾数量（“截尾”， $N=49$ ，29.7%）和总数（“总数”， $N=165$ ，100%）。如果个案已经被剔除（不在上面的例子中），则可以从“未使用的个案”一栏查询原因（例如缺失值或者负数）。

组块 0：初始组块

“组块 0：初始组块”意味着，首先建立零模型或者常数模型（PLK 的系数等于零，也就是“没有”PLK 的模型），为此测算第一个基准量。也就是说，“组块 0”反映了一个模型在 PLK 进入之前的第一个参数。

其他 COXREG 输出结果的结构取决于选择了哪种变量选择方法，例如，进入法在第一个步骤后立即停止，逐步法通常则继续计数。

模型系数似然比检验	
-2	
对数	
似然值	
1062.266	

表“模型系数似然比检验”说明了没有 PLK 的模型（常数模型或者零模型）具有什么样的性能。作为基准量，测定了-2 对数似然值为 1062.266。在下一步重新测定-2 对数似然值，并将其用于模型性能的比较。之后显示的卡方值表现了前后连续的-2 对数似然值之间的差值。

组块 1：方法 = 进入

模型系数似然比检验^{a,b}

-2 对数似然值	总 (值)			源自前一个步骤的变化			源自前一个组块的变化		
	卡方	df	显著性	卡方	df	显著性	卡方	df	显著性
1049.572	13.787	1	0.000	12.694	1	0.000	12.694	1	0.000

a.初始组块编号 0，先验对数似然值函数，-2 对数似然值：1062.266

b.从组块 1 开始。方法 = 进入

在“组块 1：方法 = 进入”后面输出“模型系数似然比检验”表。“组块 1”反映了一个模型在 PLK 进入之后的参数。

“-2 对数似然值”一行反映了在 PLK 进入模型后的数值（1049.572）。这个值与组块 0 的 -2 对数似然值（-2log likelihood，-2LL）的差值是由模型中的 PLK 的效应造成的，从而使“源自前一个步骤的变化”一行的卡方值为 12.694。

首先进行一次全面的假设检验，在“总值”一行得出一个显著性结果（ $p=0.000$ ，“显著性”），和“源自前一个步骤/组块的变化”两行一样。

如果采用逐步法一次性给出多个变量（在本例中不是这样），则“源自前一个组块的变化”一行含有逐步累积的卡方值，“源自前一个步骤的变化”一行含有自由度。如果只有一个变量，则这两行的参数完全一样。如果对数似然值统计量和之后的 Wald 检验在结果上不一致，则在有疑问的情况下应优先选用对数似然值统计量。

排除零假设意味着，应排除这个假设：任何一个进入模型的变量对模型都没有影响。在本例中，成功地排除了这个零假设：PLK 对生存时间没有影响。由于显著性低于 0.05，因此可以得出结论：PLK 在统计学上对模型有显著影响。

备注：这些差值可能有小幅变化，因为 SPSS 内部是用更为精确的数值进行计算的，所以输出结果可能由于四舍五入而有所不同。

方程中的变量

	B	SE	Wald	df	显著性	Exp(B)	Exp(B)的 95%置信区间	
							下限	上限
Plk	0.030	0.008	13.795	1	0.000	1.030	1.014	1.046

表“方程中的变量”反映了所测定模型的变量和参数。针对进入模型的协变量 PLK 的效应，给出了不同的参数。“B”是估计的、非标准化回归系数（0.030），为此显示 B 的标准误差（对于分类变量输出一个有少许不同的表格，参见第 4.7.4 节）。如果显著性低于 0.05，则可以由此得出结论：协变量对模型有影响。

对于回归系数 B，正负号和数值都是很重要的。正数的系数（例如 $PLK=0.030$ ）意味着，协变量（PLK）提高了发生目标事件的风险率（或者降低了生存概率）；负数的系数则降低了发生目标事件的风险率（或者提高了生存概率）。

需要注意的是，SPSS 输出非标准化系数；因此对这种非标准化系数的估计无论如何是不

可靠的。 $\text{Exp}(B)$ 在此可以起到指明方向的作用（见下文）： $\text{Exp}(B)$ 的数值与 1 的偏差越大，相对来看回归系数也就越大。 B 和 $\text{EXP}(B)$ 相互之间呈下列关系： $\text{EXP}(B)=1$ 相当于 $B=0$ ，即变量无影响； $\text{EXP}(B)>1$ 相当于 $B > 0$ （影响减小，见上文）； $\text{EXP}(B)<1$ 相当于 $B < 0$ （影响增大，见上文）。 B 除以其标准误差的平方就得出了 Wald 统计量 (B/SE^2)。由于 Wald 统计量 (13.795，参见“Wald”) 达到了统计显著性 (0.000，参见“显著性”) (小于 0.05)，因此 PLK 对模型是有作用的。

$\text{Exp}(B)$ 表明了预测变量 PLK 升高一个单位的风险率的预测变化。在 Cox 回归中，“ $\text{Exp}(B)$ ”就是风险比（不能与比值比混淆）。 $\text{Exp}(B)$ 表明了当解释（在这里是定量变量）变量升高一个单位时，危险函数上升或者下降几倍。始终相对于 1 解释 $\text{Exp}(B)$ 。 $\text{Exp}(B)>1$ 表示 β 系数为正数，即协变量越大，发生目标事件的风险越大。 $\text{Exp}(B)<1$ 表示 β 系数为负数，即协变量越大，发生目标事件的风险越小。风险比是一个常数，可以在一个组内部或者几个组之间予以解释。对于 $\text{Exp}(B)$ 而言，重要的是所调出的置信区间是否包含 1；如果包含了 1，则协变量没有值得一提的效应。置信区间的数值超过 1 越多，相应变量的影响也就越大。可以十分简便地通过 $\text{Exp}(B)$ 解读非标准化定量协变量的相对意义（ B 是非标准化的，可能有很大的误导作用）。通常，只解释显著性协变量的 $\text{Exp}(B)$ 。概率可以解释为根据其他协变量做了调整。对于定量变量，仅仅说明 $\text{Exp}(B)$ 是不够的。

上面的例子没有研究各组之间的比较；因此，只能在一组内部对风险率进行解释。以风险率具有比例性的假设成立为前提条件，例如，上面的例子对于一组内部的比较意味着：如果协变量升高一个单位，则风险率升高到 1.03，即升高 3%。相反，如果协变量 PLK 升高 5 个单位 ($1.035 = 1.159$)，则根据因子 1.159 将风险率升高，升幅为 15.9%。

在各组之间进行比较时，可以将风险比解释为两个恒定的风险率之间的比例（前提条件是风险率具有比例性的假设成立），也就是协变量效应的恒定比例。换言之（假设上面测定的风险比是基于各组之间的比较）：对于这个个案，可以说在其中一个组中的 PLK 风险与另一个组的相比始终是 1.03:1。在一个 1.0 编码只有两组的个案中，风险比可以解释为相对风险。

在给出 $\text{Exp}(B)$ 时，应始终同时给出相应比例的单位。

Allison (2001, 117) 建议，对于定量协变量使用公式 $(\text{Exp}(B)-1)*100$ 作为解释的辅助手段。得出的数值表示当协变量变化一个单位时估计到的百分比变化。对于 PLK 而言，通过这个方法 $((1.030-1)*100)$ 得出的数值为 3%，也就是说，PLK 每增加一个单位，生存概率降低 3%。但是，在（草率地）判定这两者之间具有实质性关联之前，还应检查 $\text{Exp}(B)$ 的置信区间。因为尽管 $\text{Exp}(B)$ 的置信区间包括 1，但是有 95%的概率可以认为，PLK 没有有效的效应。

总的来说，显著性检验、点估计 ($\text{Exp}(B)$) 和置信区间可以用作判断的辅助手段。如何使用这些判断辅助手段？置信区间对零假设检验起到补充作用，在实际应用回归分析时甚至优先使用置信区间，因为置信区间是不受取样范围影响的。如果一个置信区间不包括数值 1，则表明这是一个统计上具有显著性的结果，是不受取样范围影响的。如果置信区间和零假设的结果相互矛盾，则通常优先采用基于置信区间的结果。

生存表

时间	累积风险率的基本数值	协变量的平均值		
		生存分析	SE	累积风险率
0	0.015	0.971	0.013	0.029
1	0.072	0.872	0.024	0.137
2	0.090	0.842	0.027	0.172
3	0.113	0.807	0.029	0.215
4	0.120	0.795	0.030	0.229
5	0.140	0.765	0.032	0.268
6	0.144	0.759	0.032	0.275
7	0.174	0.717	0.034	0.333
8	0.179	0.711	0.034	0.341
9	0.193	0.693	0.035	0.367
10	0.226	0.650	0.036	0.431
11	0.236	0.638	0.037	0.450
12	0.251	0.619	0.037	0.479
13	0.284	0.582	0.038	0.541
14	0.290	0.576	0.038	0.552
15	0.301	0.563	0.038	0.574
16	0.307	0.557	0.038	0.586
17	0.338	0.525	0.038	0.645
18	0.351	0.512	0.038	0.670
19	0.365	0.499	0.039	0.696
20	0.372	0.492	0.039	0.709
21	0.379	0.486	0.039	0.723
22	0.408	0.459	0.039	0.778
23	0.416	0.453	0.039	0.793
24	0.423	0.446	0.039	0.807
25	0.439	0.433	0.038	0.837
26	0.456	0.419	0.038	0.869
30	0.464	0.413	0.038	0.885
33	0.481	0.400	0.038	0.917
34	0.499	0.386	0.038	0.952
35	0.528	0.366	0.038	1.006
36	0.537	0.359	0.038	1.024
39	0.547	0.352	0.037	1.043
41	0.557	0.346	0.037	1.062
43	0.568	0.339	0.037	1.082
44	0.578	0.332	0.037	1.102
49	0.589	0.325	0.037	1.123
53	0.603	0.317	0.037	1.150
56	0.620	0.307	0.037	1.182

续表

时间	累积风险率的基本数值	协变量的平均值		
		生存分析	SE	累积风险率
62	0.642	0.294	0.037	1.223
63	0.667	0.280	0.038	1.272
66	0.731	0.248	0.039	1.394
80	0.801	0.217	0.042	1.527

“生存表”反映了“累积风险率的基本数值”，也就是所给出的协变量 **PLK** 的平均值（在“在协变量的平均值上”一项下）、平均风险率（“生存分析”）、其标准误差（**SE**）和累积风险率。“累积风险率的基本数值”一列含有假设个体的风险估计值，前提条件是这个个体的所有协变量都等于零；在给出的时间点、不受协变量效应影响的对风险率进行估计。如果由于协变量的平均值等于零（否则就编程给出协变量的平均值），而具有标准化的数值变量，或者如果所有协变量是分类协变量，则这一列的数据特别有用。“累积风险率的基本数值”一列反映了没有协变量、而是只注明时间的 **Cox** 模型的风险率。相应的，风险率随着时间累积增长。如果所有协变量都是分类协变量，则在类别 0 中测定这些个案的风险率。

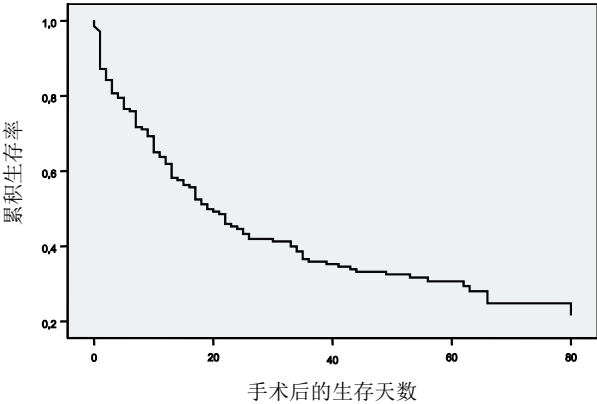
“在协变量的平均值”一项下的参数是对具有“平均”协变量的假设个体的估计量，这对于一个假设个案与实际患者的比较可能是有用的。如果有分类协变量，则这个比较的可用性可能就受到了限制。“生存分析”一列含有在相应的时间点还没有达到目标事件的患者的估计比例（包括相应的标准误差“**SE**”）。例如，在一项癌症调查中，在这个时间点患者仍旧生存。例如，在调查的第 3 天，有 80% 的调查参与者还没有出现目标事件；在调查的第 33 天，只有 40% 的调查参与者还没有出现目标事件。“累积风险率”一列给出了出现目标事件的风险或者反向风险。危险函数相当于生存函数的负对数，因此只是（预测的）生存概率数据的一种数学表现形式。

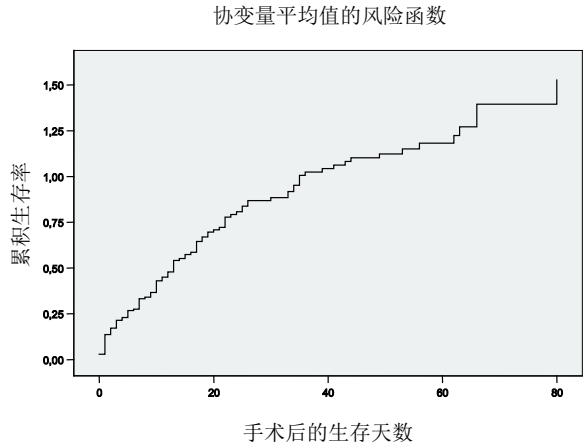
协变量平均值

	平均值
Plk	21.733

表“协变量平均值”反应了协变量平均值（例如在这里是 **PLK**），它是测定累积风险率的基础，针对的是“生存表”的“在协变量的平均值”一列。

协变量平均值的生存函数

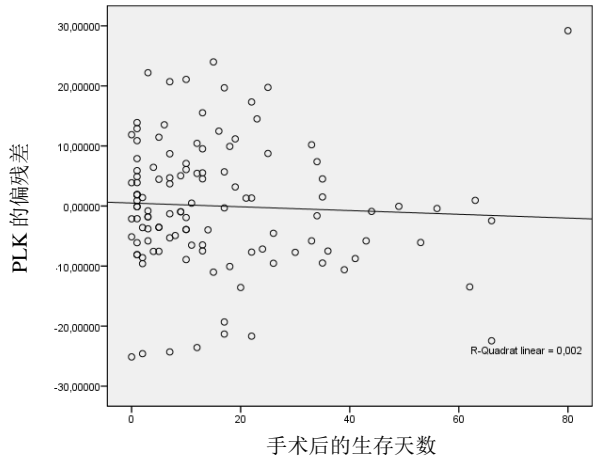




累积生存函数和累积危险函数的趋势是相反的。在生存逐渐累积时，生存概率随之下降，从而不会出现目标事件相反；在风险逐渐累积时，风险随之升高，从而出现目标事件。从数学角度看，这两个函数是相互对应的，生存函数的负对数就是危险函数。如果在手术后生存时间的某些区间（例如从第 60 天到第 80 天），只有少数患者仍停留在这项调查研究范围内，则根据患者的人数不同，函数曲线可能变得有些“粗略”。

图“累积生存率”（上方图）展示了由模型预测的“平均水平”患者的生存时间。 Z 轴表示到出现目标事件为止的时间， Y 轴表示“平均的”累积生存率。函数曲线上的每个点都表示累积生存概率，也就是一个“平均水平”患者在这个点之后仍然生存的概率。

图“累积风险率”（下方图）展示了由模型预测的“平均水平”患者的风险。 X 轴表示直到发生目标事件为止的时间， Y 轴表示平均累积风险率（相当于生存概率的负对数）。函数曲线上的每个点都表示累积风险，也就是一个“平均水平”患者出现目标事件的风险。



最后输出的散点图（Schoenfeld 残差图）在 X 轴上反映了自手术之后的生存时间，在 Y 轴上反映了 PLK 的偏残差。协变量的偏残差是观察值与基于模型的期望值之间的偏差。这个图以图形的方式检验了风险随着时间变化具有比例性的前提条件是否成立。Cox 模型的其中一个

假设（准确来说：表示时间不相依协变量的模型 1）是定距协变量（例如，这里的 PLK），与时间不相依。因此，应始终检验协变量是否与时间不相依，如果这个假设不成立，则必须计算表示时间相依协变量的模型 2。如果每次截取的协变量的效应随着时间的推移具有比例性，则残差每次平均都会得出数值 0。因此，图中的点云不应具有任何呈系统性的趋势。绘出的回归曲线应尽量接近于零。之后记录的（线性）回归方程与零不具有明显误差，因此说明每次截取的协变量 PLK 的效应随着时间的推移具有比例性。对于利用模型 1 的这个分析而言，随着时间变化风险具有比例性的前提条件得到了满足。可以说，这个检验的可靠性取决于模型中未截尾个案的数量。只针对未截尾个案测定了偏残差。一个模型中所含有的未截尾个案越多，这个借助图形的前提条件检验就越可靠。在关于 Cox 回归一章的结尾部分介绍了其他前提条件检验，以及针对没有满足数据的各种前提条件的个案可以采取哪些措施（参见第 4.7.6 节和 4.7.7 节）。

下面的几节展示了如何将 Cox 回归的基本模型扩展增加分类协变量（分层变量，Strata）或者时间相依协变量。

4.7.3 带有二元协变量的 Cox 回归（ $k=2$ ）

提出的问题

这个分析的目的是检验某种药物（实验药物 vs. 安慰剂）是否以及如何对 $N=165$ 个患者的手术后生存时间的变化趋势产生影响（参见第 4.6.6 节）。在本例中，协变量 MED 是二元尺度的。选择变量 MED 的最后一个类别作为参考类别。

由于与前面已经介绍的示例有一些重叠，因此下面的图示只集中讲解这个分析的特殊之处。为了集中讲解关键部分，已知的界面选择路径、SPSS 语句中记录的设置、完全相同的或者至少十分类似的输出结果在这里不再阐述或者解释。

在 SPSS 程序主界面选择以下菜单项：分析 → 生存函数 → Cox 回归...

将变量“UEBERLEB（生存）”拖入“时间”栏。将变量“UESTATUS（状态）”拖入“状态”栏。在“定义事件”一栏下给出将要发生的目标事件的编码，例如“0”。将变量 MED 拖入“协变量”栏。在“方法”一项下选定“进入”。

子窗口“定类...”：将 MED 从选择窗口拖入“分类变量”。在“修改对比方法”一项下调整对比方法“指示符”。单击“修改”按钮，应用这个对比方法。在“参考类别”一项下确定与所有其他水平做比较的参考类别。单击“最后”按钮，使用变量 ST 的最后一个类别作为参考类别。单击“继续”按钮。

备注：在“指示符”一项中，对比指的是是否从属于某个类别。在对比矩阵中，参考类别显示为带有零的横行。在第 4.7.9 节中解释了其他的对比方法。应事先指出的是，如果选择 SPECIAL（“特别”）命令，用户就能使用可以自己定义的对比方法。

子窗口“绘图”：选定选项“生存函数”和“危险函数”。不进行下一步的设置。单击“继续”按钮。

单击“确定”按钮开始计算。

语句:

```
COXREG
ueberleb /STATUS=uestatus(0)
/PATTERN BY med
/CONTRAST (med)=Indicator
/METHOD=ENTER med
/PLOT SURVIVAL HAZARD
/CRITERIA=PIN(.05) POUT(.10) ITERATE(20)
/PRINT=CI(95) BASELINE.
```

备注：利用 **PATTERN BY** 命令，确定所调出图形的成组变量。例如，在这些图中，针对 **MED** 的每个类别都输出了一条单独的线条。**BY** 命令后的变量必须是定类尺度的。在单独的 **PATTERN BY** 命令行中，只要在 **/METHOD=** 或者 **/CONTRAST=** 后面已经做了指定，就可以给出其他的定类尺度协变量。如果这个模型含有时间相依协变量，则不能使用 **PATTERN**。

通过 **CONTRAST** 命令，确定对所给出定类尺度协变量的各个变量类别进行比较，以及这些比较（对比）应是哪种类型。对于协变量 **MED**，进行了“指示符”对比方法，第 4.7.9 节归纳了其和其他对比方法的相关内容。应事先指出的是，通过带有 **SPECIAL** 的语句，用户能使用可以自己编程设定的对比方法。可以给出多个 **CONTRAST** 行，前提条件是至少在 **/METHOD** 一项下给出了定类尺度协变量。

输出结果

Cox 回归

表“个案处理评估”与第 4.7.2 节的表格完全相同，因此不再赘述。

分类变量的编码 ^b			
		频率	(1)
MED ^a	1=实验药物	33	1
	2=安慰剂	132	0

- a. 指示参数的编码。
- b. 分类变量：**MED**（药物）。

表“分类变量的编码”反映了定类尺度协变量的内部编码。如果没有给出定类尺度协变量，则无法明确地解释（尤其是对于分类协变量）所输出的参数（如风险比）。在本例中，所测定的统计量针对的是与参考类别 **MED=2**（“安慰剂”，哑变量编码=0）做比较的事件“实验药物”（**MED=1**，哑变量编码=1）。如果事件针对的是与“实验药物”（参考类别）做比较的“安慰剂”，则应在开头设定另一个参考类别。例如，通过从数据集中给出相应变量的准确编码值，或者通过将变量重新编码（如通过 **RECODE** 命令）就可以实现这一点。

在表下面，通过“指示参数”指出了对比的类型：例如，如果设定了方法“误差”（**DEVIATION**），在这里就应出现“误差参数”。最后给出协变量的名称和标签。

在“组块 0：初始组块”中的表“模型系数似然比检验”与第 4.7.2 节的完全相同，因此

不再赘述。

组块 1：方法 = 进入

模型系数似然比检验^{a,b}

-2 对数似然值	总 (值)			源自前一个步骤的变化			源自前一个组块的变化		
	卡方	df	显著性	卡方	df	显著性	卡方	df	显著性
1048.786	11.539	1	0.001	13.479	1	0.000	13.479	1	0.000

a. 初始组块编号 0，先验对数似然值函数，-2 对数似然值：1062.266。

b. 从组块 1 开始。方法 = 进入。

在“组块 1：方法 = 进入”后面，表“模型系数似然比检验”反映了变量 MED 进入模型之后的模型参数。

“-2 对数似然值”一行反映了在 PLK 进入模型后的数值（1048.786）。与组块 0 中的-2 对数似然值的差异是由变量 MED 进入模型造成的，导致“变化”几列中的卡方值等于 13.479。

由于“显著性”（ $p=0.000$ ）的数值低于 0.05，因此可以得出结论：变量 MED 对于模型具有统计学意义上的影响。

方程中的变量

	B	SE	Wald	df	显著性	Exp(B)	Exp(B)的 95%置信区间	
							下限	上限
med	-0.937	0.286	10.743	1	0.001	0.392	0.224	0.686

表“方程中的变量”反映了所测定模型的变量和参数。对于具有超过两个层的分类变量，输出了一个稍有不同的表格（参见第 4.7.4 节）。针对进入模型的二元协变量 MED 的效应，给出了不同的参数。“B”是估计的非标准化回归系数（-0.937），对此显示 B 的标准误差。Wald 统计量（10.743，关于其测定参见第 4.7.2 节）得出了统计上的显著性（ $p=0.001$ ）。因此，变量 MED 对于模型是有用的。二元变量 MED 对于手术后生存时间的变化趋势具有统计学意义上的效应。Exp(B)表明了预测变量 MED 升高一个单位时预测的风险率变化（关于 Exp(B)与 B 的关系参见第 4.7.2 节）。Exp(B)的置信区间排除了 1，可以认为，变量 MED 施加了统计学上有意义的效应。对于二元尺度变量的 Exp(B)的解读，与定量变量的有所不同（参见第 4.7.2 节）。

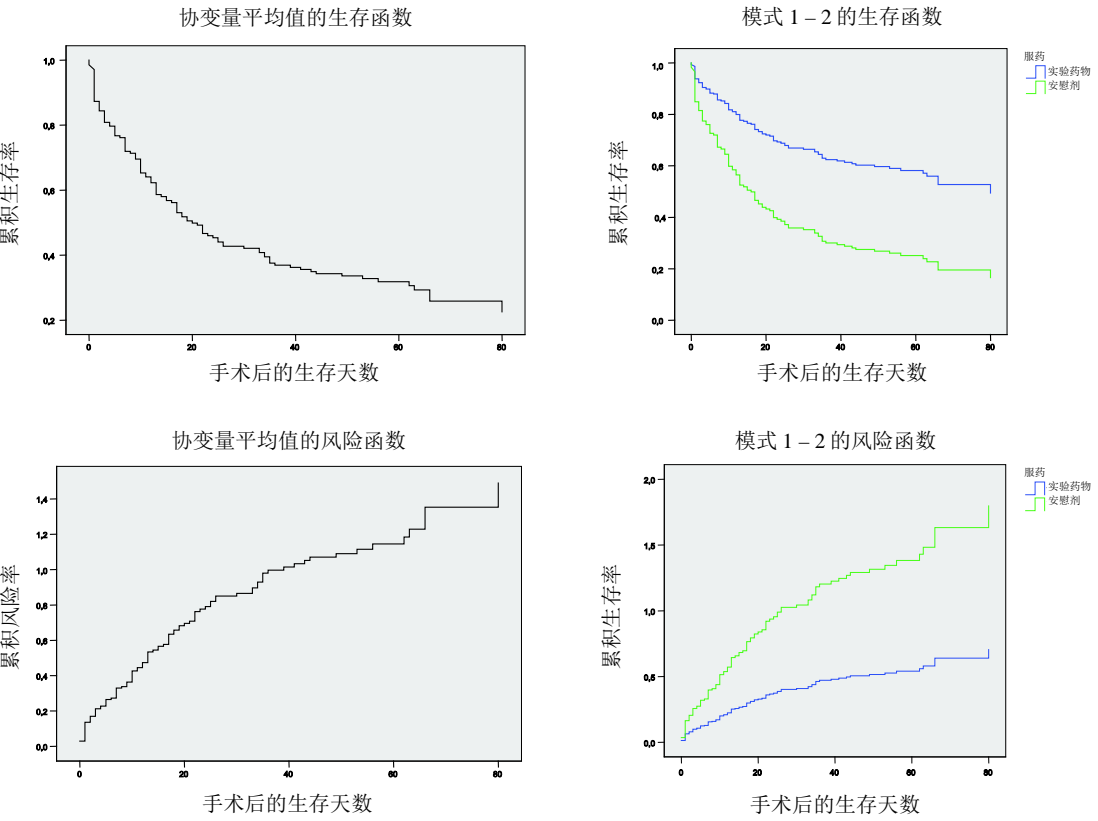
由于 MED 是一种二元分类变量，因此可以直接解读实验药物和安慰剂两个组的风险率。给出的检验方向“实验药物”vs.“安慰剂”（“参考类别”，哑变量编码=0）表明，实验药物一组的风险是安慰剂一组的 0.392 倍。对于相反的检验方向“安慰剂”vs.“实验药物”，根据呈倒数的 Exp(B)值（ $1 / 0.392 \approx 2.55$ ）得出，安慰剂一组的风险大约是实验药物一组的 2.5 倍。

在本例中，“生存表”的结构和内容都与第 4.7.2 节的基本相同（也可参见下面给出的单变量生存曲线（左边）），因此不再赘述。

协变量平均值和模式值

	平均值	模式	
		1	2
med	0.200	1.000	0.000

表“协变量平均值和模式值”反应了协变量数值（例如在这里是 MED），它是测定累积风险率的基础，针对的是“生存表”的“在协变量的平均值上”一列。“模式”一列反映了所给出分类协变量（例如 MED）的各种变量类别，这些信息重新反映在分组的直线图中（参见标题）。



这些图反映了由模型预测的“平均水平”患者生存时间或者预测风险（参见第 4.7.2 节）。右边的图是根据所给出二元尺度协变量的变量类别分组的。针对所给出二元尺度协变量的平均值输出了左边的图，但是这个意义非常有限。很明显可以看出，“实验药物”组的风险比“安慰剂”组低。Exp(B)代表了两条线之间的距离。从“实验药物”组的角度来看，风险比较低；从“安慰剂”组的角度来看，风险比较高。

4.7.4 带有分类协变量的 Cox 回归 (k>2)

提出的问题

这个分析的目的是，检验四个组 (k=4) 之间 N=165 个患者的术后生存时间变化趋势是否有区别。根据病情的严重程度，将这些患者分为 I（轻微）至 IV（严重）四个组。选择变量

ST 的第一个类别作为参考类别。在本例，协变量 ST 是四级尺度的。

由于与前面已经介绍的示例有一些重叠，因此下面的图示只集中于这个分析的特殊之处。为了集中于关键部分，已知的界面选择路径、SPSS 语句中记录的设置、完全相同的或者至少十分类似的输出结果在这里不再阐述或者解释。

在 SPSS 程序主界面选择以下菜单项：分析 → 生存函数 → Cox 回归...
将变量 ST 拖入“协变量”栏。在“方法”一项下选定“进入”。

子窗口“定类...”：将 ST 从选择窗口拖入“分类变量”。在“参考类别”一项下确定类别“第一个”，ST 的所有其他类别（ST=2，3，4）与这个类别（ST=1）相比较。单击“修改”按钮，应用这个设置。单击“继续”按钮。

备注：在第 4.7.9 节中解释了其他的对比方法。应事先指出的是，如果选择 SPECIAL（“特别”）命令，则用户就能使用可以自己定义的对比如方法。

单击“确定”按钮，开始计算。

语句：

```
COXREG
  ueberleb  /STATUS=uestatus(0)
  /PATTERN BY st
  /CONTRAST (st)=Indicator(1)
  /METHOD=ENTER st
  /PLOT SURVIVAL HAZARD
  /PRINT=CI(95) BASELINE
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) .
```

备注：通过 CONTRAST 命令，在括号内给出了参考类别。在本例中，变量 ST 的所有其他变量类别与变量类别 1 进行比较。

输出结果

Cox 回归

表“个案处理评估”与第 4.7.2 节的完全相同，因此不再赘述。

分类变量的编码^b

		频率	(1)	(2)	(3)
st ^a	1=I	70	0	0	0
	2=II	33	1	0	0
	3=III	54	0	1	0
	4=IV	8	0	0	1

a. 指标参数的编码。
b. 分类变量：st（阶段）。

表“分类变量的编码”反映了定类尺度协变量 ST 的内部编码。根据给定的检验方向，将变量类别 2 至 4 与 ST=1（参考类别）进行比较。因此，所测定的统计量始终针对的是参考类

别 ST=1 (“I”，哑变量编码：[0 0 0])。如果这些事件针对的是其他变量类别，则应设置另一个参考类别或者将变量重新编码。

在表格下面，通过“指示参数”指出了对比的类型。例如，如果设定了方法“误差”(DEVIATION)，在这里就应出现“误差参数”。最后给出协变量的名称和标签。由于变量 ST 具有 n 个等级，因此定义了 $n-1$ 个哑变量，将这些哑变量纳入模型方程，也就是说，单独地和针对整体影响地测定其系数。

在“组块 0：初始组块”中的表“模型系数似然比检验”与第 4.7.2 节和 4.7.3 节的完全相同，因此不再赘述。

组块 1：方法 = 进入

模型系数似然比检验^{a,b}

-2 对数似然值	总 (值)			源自上一个步骤的变化			源自前一个组块的变化		
	卡方	df	显著性	卡方	df	显著性	卡方	df	显著性
1044.247	18.063	3	0.000	18.019	3	0.000	18.019	3	0.000

a. 初始组块编号 0，先验对数似然值函数，-2 对数似然值：1062.266。

b. 从组块 1 开始。方法 = 进入。

在“组块 1：方法 = 进入”后面，表“模型系数似然比检验”反映了变量 MED 进入模型之后的模型参数。“-2 对数似然值”一行反映了变量 ST 进入模型后的数值 (1044.247)。变量 ST 进入模型使得与组块 0 中的 -2 对数似然值产生差异，得出卡方值为 18.019。

由于“显著性”(p=0.000)的数值低于 0.05，因此可以得出结论：变量 ST 对于模型具有统计学意义上的影响。

方程中的变量

	B	SE	Wald	df	显著性	Exp(B)	Exp(B)的 95%置信区间	
							下限	上限
st			17.124	3	0.001			
st(1)	0.923	0.244	14.327	1	0.000	2.517	1.561	4.059
st(2)	0.750	0.227	10.925	1	0.001	2.116	1.357	3.300
st(3)	0.595	0.441	1.818	1	0.178	1.812	0.764	4.301

对于具有超过两个层的分类变量，表“方程中的变量”与前面示例的表格（例如第 4.7.2 节和 4.7.3 节）稍有不同。针对进入模型的多元协变量 ST 的效应，给出了 $k-1$ 个参数。变量 ST 的第一行执行了变量的总检验，检验结果具有显著性 (Wald=17.125, p=0.001)，因此，ST 对于模型是有用的。分类变量 ST 对于 165 位患者术后生存时间的变化趋势具有统计学意义上的影响。由于 ST 是一种多元分类变量，因此风险比看起来与二元协变量（参见 4.7.3 节）的不同。下面几行则是与希望的参考类别 (ST1) 的比较：

“st(1)” 相当于 ST=2 vs. ST=1: exp(B)=2.517

“st(2)” 相当于 ST=3 vs. ST=1: exp(B)=2.116

“st(3)” 相当于 ST=4 vs. ST=1: exp(B)=1.812

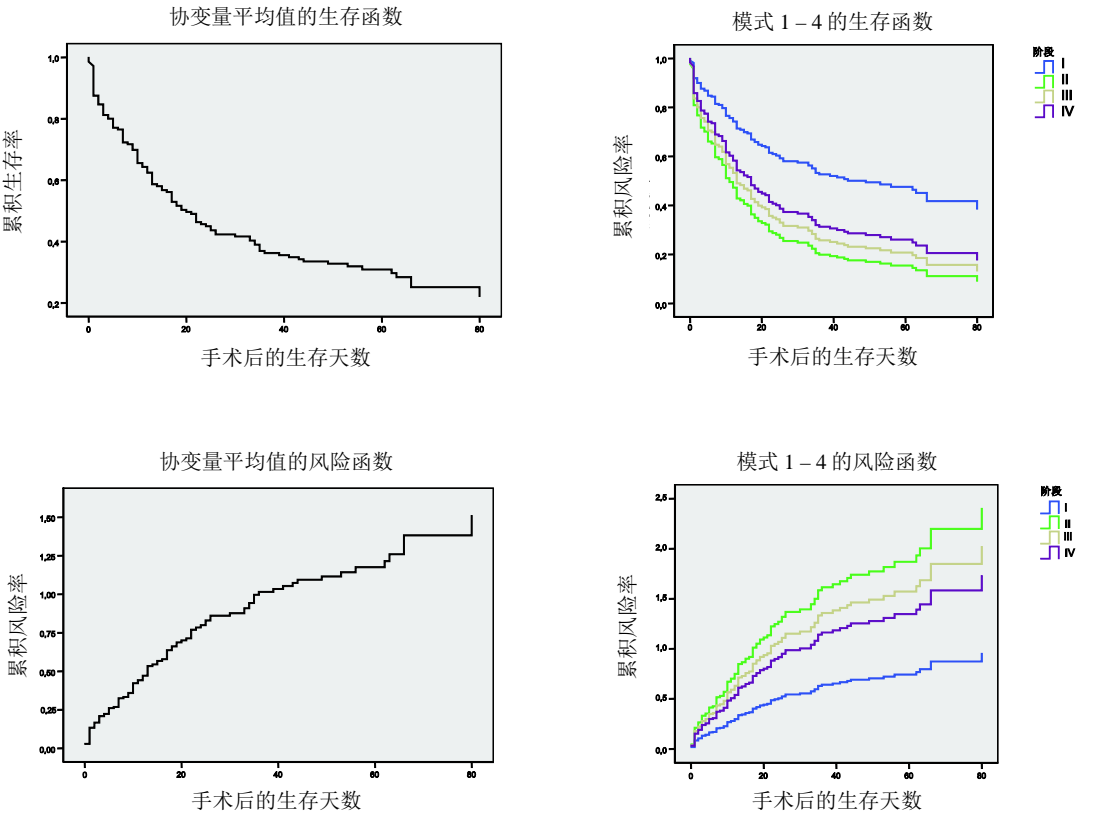
例如，检验方向 ST=2 vs. ST=1（“参考类别”）得出，ST=2 组的风险是参考类别 ST=1 的 2.517 倍。

Exp(B)的所有置信区间（除了最后一个比较之外）排除了 1，可以认为，在这里出现了统计学意义上的效应。最后一个比较，即 ST=4 组与参考类别的比较只基于 N=8 个个案；也就是个案数量比其他的比较少了很多倍（参见表“分类变量的编码”）。由于个案数量少了很多，因此不应草率地驳回对于“st（3）”的调查结果。从下面的图可以看出，分类变量 ST 的效应是建立在这一点的基础上：ST=1 这组的变化趋势有利得多。

在本例中，“生存表”的结构和内容都与第 4.7.2 节的表格基本相同（也可参见下面给出的单变量生存曲线（左边）），因此不再赘述。

协变量平均值和模式值					
	平均值	模式			
		1	2	3	4
st(1)	0.200	0.000	1.000	0.000	0.000
st(2)	0.327	0.000	0.000	1.000	0.000
st(3)	0.048	0.000	0.000	0.000	1.000

表“协变量平均值和模式值”反映了协变量数值（例如在这里是 ST），它是测定累积风险率的基础，针对的是“生存表”的“在协变量的平均值”一列。“模式”一列反映了分类协变量 ST 的变量类别，这些信息重新反映在分组的直线图中（参见标题）。



这些图反映了由模型预测的“平均水平”患者生存时间或者风险（参见第 4.7.2 节）。右边的图则是根据分类协变量 ST 的变量类别分组的。针对 ST 的平均值输出了左边的图，但是这个意义非常有限。很明显可以看出，“I”组的风险比其他三组低。Exp(B)代表了“I”组的直线和其他三组的直线之间的距离。患者病情的严重程度对于是否生存具有显著的影响。从这些图可以看出，分类变量 ST 的效应是建立在这一点的基础上：ST=1 这组的变化趋势有利得多。

4.7.5 针对交互作用的 Cox 回归

只要模型含有超过一个主因素，就可以对交互作用进行建模。如果模型只有一个主因素，则模拟了一个未明确表达的假设：除了这个唯一的主因素之外，任何其他因素都不会对生存时间有影响。相反，如果两个或者多个主因素之间有交互作用，则模拟了这个假设：各个主因素之间具有交互作用，并且对生存时间有影响（Klein&Moeschberger, 2003, 250-253）。

Kleinbaum & Klein（2005, 26-27）指出了与第三方变量的交互作用可能产生的效应。Allison（2001, 14）建议将所有可以说明截尾规模的因素纳入一个模型。

只要可以假设，一旦利用第二个因素在一个共同的模型里指定了第一个因素的风险比，则第一个因素的风险比就发生变化，那么这两个因素之间的交互作用对于我们所研究的问题就是有用的。交互作用既可以用定量协变量来建模（例 I），也可以用定类尺度协变量（例 II 和 III）来建模。

例 I：两个定量变量

提出的问题

以下所做的分析是为了检验，两个定量影响变量 T_GROESSE（肿瘤尺寸）和 LYM_KNOT（淋巴结数量）是否作为主效应或者交互作用对于 N=171 位患者的生存时间有影响。

图示主要集中在当前分析实例的关键部分。已知的界面选择路径、SPSS 语句中记录的设置、完全相同的或者至少十分类似的输出结果在这里不再阐述或者解释。

在 SPSS 程序主界面选择以下菜单项：分析 → 生存函数 → Cox 回归...

将两个定量变量 T_GROESSE 和 LYM_KNOT 拖入“协变量”一栏。再次选定在变量列表左边的这两个变量，用下面的按键“>b”同时将其成对地拖入“协变量”一栏。在“方法”一项下选定“进入”。

单击“确定”按钮开始计算。

语句：

```
COXREG
    ueberleb
    /STATUS=uestatus(1)
```

```
/METHOD=ENTER lym_knot t_groesse lym_knot*t_groesse
/PLOT SURVIVAL HAZARD
/PRINT=CI(95) BASELINE
/CRITERIA=PIN(.05) POUT(.10) ITERATE(20) .
```

备注：在/METHOD 一项下，利用 ENTER 命令调出“进入”法。在 ENTER 后面给出想要的效应，在本例中就是两个定距变量 LYM_KNOT 和 T_GROESSE，以及两者的交互作用。

输出结果

Cox 回归

个案处理评估			
		N	百分比
分析时的可用个案 未使用的个案	事件 ^a	66	38.6%
	已截尾总共缺失了数值	86	50.3%
	的个案	152	88.9%
	负数时间的个案	19	11.1%
	一个班次中出现最早事	0	0.0%
	件之前的截尾个案	0	0.0%
	总数	19	11.1%
		171	100.0%

a. 因变量：生存时间（月）

不再阐述表“个案处理评估”。

组块 0：初始组块

模型系数似然比检验
-2 对数似然值
572.873

不再阐述表“模型系数似然比检验”。

组块 1：方法 = 进入

模型系数似然比检验 ^{a,b}									
-2 对数似然值	总（值）			源自上一个步骤的变化			源自前一个组块的变化		
	卡方	df	显著性	卡方	df	显著性	卡方	df	显著性
496.925	104.615	3	0.000	75.947	3	0.000	75.947	3	0.000

a. 初始组块编号 0，先验对数似然值函数，-2 对数似然值：572.873。

b. 从组块 1 开始。方法 = 进入。

不再阐述表“模型系数似然比检验”。

不再阐述表“生存表（没有显示）”。

方程中的变量

	B	SE	Wald	df	显著性	Exp(B)	Exp(B)的 95%置信区间	
							下限	上限
lym_knot	0.191	0.048	15.917	1	0.000	2.517	1.102	1.330
t_grpesse	0.723	0.082	77.429	1	0.000	2.116	1.754	2.420
lym_knot*t_grpesse	-0.051	0.017	9.498	1	0.002	1.812	0.920	0.982

两个定距变量 LYM_KNOT 和 T_GROESSE 以及两者的交互作用达到了统计显著性。围绕着 Exp(B)的置信区间，不包括 1。对于这些协变量产生了下列效应：

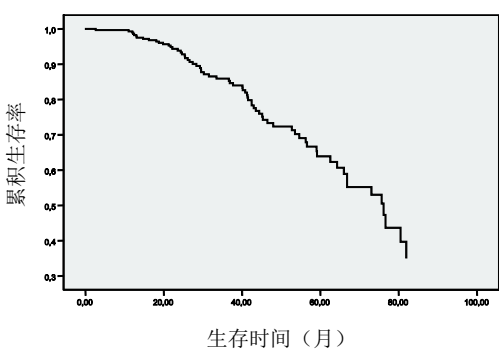
- LYM_KNOT: $[(1.211-1)*100]=21.1\%$ 。LYM_KNOT 每增加一个单位，生存概率升高 21.1%。
- T_GROESSE: $[(2.060-1)*100]=106\%$ 。T_GROESSE 每增加一个单位，生存概率升高 106%。
- LYM_KNOT*T_GROESSE: $[0.950-1)*100] = -5\%$ 。LYM_KNOT 和 T_GROESSE 之间的交互作用每增加一个单位，生存概率降低 5%。

协变量平均值

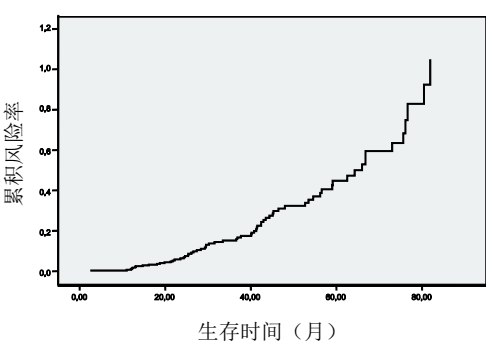
	平均值
lym_knot	1.092
t_grpesse	1.338
lym_knot*t_grpesse	2.492

不再阐述表“协变量平均值”。

协变量平均值的生存函数



协变量平均值的风险函数



这些图反映了由模型预测的“平均水平”患者生存时间或者预测风险，这里不再赘述。

备注

生存概率随着淋巴结数量增多或者肿瘤增大而升高，这个结果看上去就是反常的。原先的预测是，生存概率随着淋巴结数量增多或者肿瘤增大而降低。但是如果更精确地审视这个结果，也就是说，根据当前的取样，生存概率是随着淋巴结数量增多或者肿瘤增大而升高的，那么生存概率升高和降低的不同就说明，这个调查研究的设计可能有问题。在构成所调查样本的大部分个案中，尽管淋巴结数量很多或者肿瘤很大，但仍然具有较长的生存时间。接下来将换一种

说法，将这个现象解释为假设因素“健壮性”并予以讨论。可以用多种类型的错误解释来讨论这个反常的结果（例如）。

- **设计方案错误（肿瘤类型）。**肿瘤的种类多种多样。在做记录时可能忽视了致命性比恶性肿瘤低得多的良性肿瘤存在。例如，现在如果只提取具有良性肿瘤的个案，那么淋巴结数量或者肿瘤大小这两个因素就不会对生存概率产生因果效应，而是只取决于时间因素（因此与时间混淆）：一个个案生存的时间越长，良性肿瘤就变得越大。但是良性肿瘤本身不是患者继续生存的原因。生存的根本原因是患者的个体健壮性。但是为了简便起见，在讨论其他方面时我们这样假设：这个调查研究的对象只包含对患者生存有不利影响的恶性肿瘤。
- **取样错误。**如果所调查个案针对的是来自于一次累积取样的各种元素，则完全有可能这次取样的各个元素在特征上有根本性不同（偏差），这就是显而易见的选择效应。因为只有当患者仍在生存时，其个案才能纳入生存分析，因此对于患者仍然生存并且具有明显的肿瘤参数的个案而言，无法完全排除患者具有显著健壮性的可能性。尤其是当健壮性很低的患者可能由于不明显的肿瘤参数已经死亡时，现在如果偶然纳入了健壮性特别强的个案，则这些患者的生存可能很少受到肿瘤参数的影响，而是更多地取决于个体的健壮性。淋巴结数量或者肿瘤大小或许与时间相关，但是不一定是生存或者死亡的（唯一）原因。相反，可以假设的是，淋巴结数量或者肿瘤大小（也）取决于患者的健壮性。一个个案的患者越健壮，他的生存时间就越长，肿瘤参数的表达可能就越明显，而且这个个案的生存概率不是肯定就受到负面影响的。

对于这个结果的解释揭示了一个无法忽略的因素“健壮性”，对此必须要进一步地进行探索，但是这个因素很可能是表明这次调查研究的设计方案出现了巨大错误。

建议进行进一步的数据探索。例如，对于调查研究设计方案的回顾性分析，这个分析的结果就与在多变量统计中一样：在分析后和分析前。

例 II：一个定量变量和一个分类变量

问题

下面分析的目的是，检验定量变量 **LYM_KNOT**（淋巴结的数量）和二元变量 **HISTGRAD**（组织学分级，变量类别 2 和 3）究竟是作为主效应还是交互作用，对 $N=171$ 个患者的生存时间产生影响。

图示主要集中在当前分析实例的关键部分。已知的界面选择路径、SPSS 语句中记录的设置、完全相同的或者至少十分类似的输出结果在这里不再阐述或者解释。

语句：

COXREG

```
ueberleb /STATUS=uestatus(1)
/PATTERN BY histgrad
/CONTRAST (histgrad)=indicator(1)
/METHOD=ENTER histgrad lym_knot histgrad*lym_knot
/PLOT SURVIVAL HAZARD
```

```
/PRINT=CI(95) BASELINE
/CRITERIA=PIN(.05) POUT(.10) ITERATE(20) .
```

输出结果：

Cox 回归

个案处理评估

		N	百分比
分析时的可用个案	事件 ^a	54	31.6%
未使用的个案	已截尾总共缺失了数值的个案	62	36.3%
	负数时间的个案	116	67.8%
	一个班次中出现最早事件之前的截尾个案	55	32.2%
	总数	0	0.0%
		0	0.0%
		55	32.2%
		171	100.0%

a. 因变量：生存时间（月）

不再阐述表“个案处理评估”。

分类变量的编码^b

	频率	(1)
histgrad ^a 2=2	68	0
3=3	48	1

a. 指示参数的编码

b. 分类变量：histgrad（组织学分级）

不再阐述表“分类变量的编码”。

组块 0：初始组块

模型系数似然比检验

-2 对数似然值
428.604

不再阐述表“模型系数似然比检验”。

组块 1：方法 = 进入

模型系数似然比检验^{a,b}

-2 对数似然值	总（值）			源自上一个步骤的变化			源自前一个组块的变化		
	卡方	df	显著性	卡方	df	显著性	卡方	df	显著性
416.841	15.356	3	0.002	11.763	3	0.008	11.763	3	0.008

a. 初始组块编号 0，先验对数似然值函数，-2 对数似然值：428.604

b. 从组块 1 开始。方法 = 进入

不再阐述“模型系数似然比检验”表。

不再阐述“生存表（没有显示）”表。

方程中的变量

	B	SE	Wald	df	显著性	Exp(B)	Exp(B)的 95%置信区间	
							下限	上限
histgrad	0.620	0.307	4.070	1	0.044	1.858	1.018	3.393
lym_knot	0.098	0.038	6.545	1	0.011	1.103	1.023	1.189
histgrad*lym_knot	-0.024	0.053	0.202	1	0.653	0.976	0.880	1.083

分类变量 HISTGRAD 和定距尺度变量 LYM_KNOT 达到统计显著性，但是没有实现其交互作用($p=0.653$)。除了在交互作用上之外，围绕着 Exp(B)的置信区间不包括 1。尽管输出的参数看起来相等，但是由于其尺度不同，因此应对其进行不同的解释。由此产生下列效应。

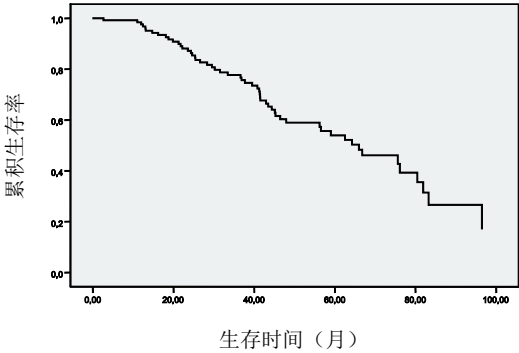
- 分类 HISTGRAD（二元）：风险率可以直接读取。检验方向“3” vs. “2”（“参考类别”，哑变量编码=0）得出，组“3”的风险是组“2”的 1.858 倍。对于（相反的）检验方向“2” vs. “3”，通过 $(1 / 1.858 \approx 0.538)$ 得出，组“2”的风险是组“3”的 0.54 倍。
- 数值：LYM_KNOT: $[(1.103-1)*100]=10.3\%$ 。LYM_KNOT 每增加一个单位，生存概率升高 10.3%。

协变量平均值和模式值

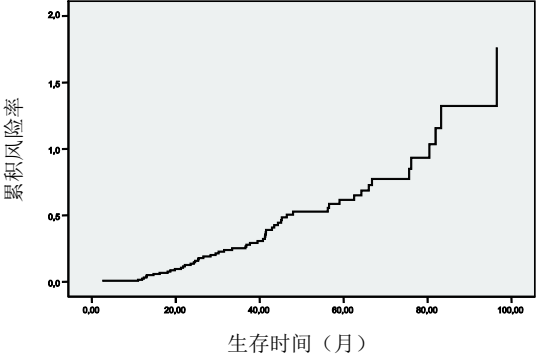
	平均值	模式	
		1	2
histgrad	0.414	0.000	1.000
lym_knot	1.371	1.371	1.371
histgrad*lym_knot	0.595	0.000	1.371

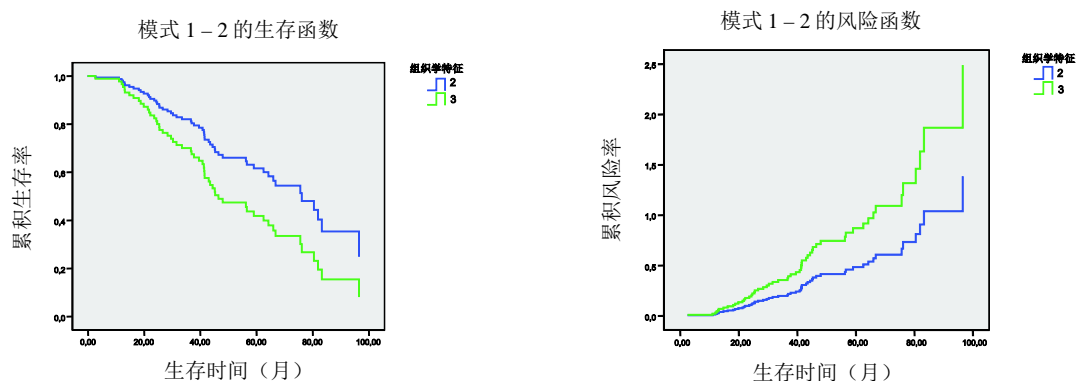
不再阐述“协变量平均值和模式值”表。

协变量平均值的生存函数



协变量平均值的风险函数





这些图反映了由模型预测的“平均水平”患者生存时间或者预测风险，这里不再赘述。

例 III：单独的“模式”生存函数

如果一个 Cox 模型含有多个分类变量，则这个问题就比较令人感兴趣，即各个分类变量（模式）的变量类别的各种组合对于生存的影响方面是否有区别。这些模式由各个分类变量的现有变量类别及其组合构成。例如，如果一个模型含有变量类别为“是”或者“否”的变量“癌症”，以及变量类别为“I”、“III”、“IV”的变量“组织学分级”，然后得出下列的组合方式。

- 模式 1：“癌症” = “是”，同时“组织学分级” = “I”
- 模式 2：“癌症” = “是”，同时“组织学分级” = “III”
- 模式 3：“癌症” = “是”，同时“组织学分级” = “IV”
- 模式 4：“癌症” = “否”，同时“组织学分级” = “I”
- 模式 5：“癌症” = “否”，同时“组织学分级” = “III”
- 模式 6：“癌症” = “否”，同时“组织学分级” = “IV”

各个分类变量现有变量类别的这些组合方式被称为“模式”，可以利用 SPSS 检验其对生存产生的影响。

提出的问题

下列分析的目的是，在考虑到定量变量 T_GROESSE 的效应的情况下，检验两个分类变量（HISTGRAD, LN_YESNO）组合起来的变量类别是否对患者各自的生存时间有影响，以及这种效应是否可以可视化。

图示主要集中在当前分析实例的关键部分。已知的界面选择路径、SPSS 语句中记录的设置、完全相同的或者至少十分类似的输出结果在这里不再阐述或者解释。

语句：

```
MEANS
  TABLES=t_groesse
  /CELLS MEAN STDDEV COUNT .
```

备注：通过 MEANS 命令确定定量变量 T_GROESSE 的平均值。T_GROESSE 的平均值为 1.3376。对于各个模式的定义需要这个数值（见下文）。

```

COXREG
  ueberleb
/STATUS=uestatus(1)
/CATEGORICAL = histgrad ln_yesno
/METHOD=ENTER t_groesse histgrad ln_yesno
/OUTFILE=TABLE( " C:\cox_surv.sav " )
/PATTERN t_groesse(1.3376) histgrad(2) ln_yesno(0)
/PATTERN t_groesse(1.3376) histgrad(3) ln_yesno(0)
/PATTERN t_groesse(1.3376) histgrad(4) ln_yesno(0)
/PATTERN t_groesse(1.3376) histgrad(2) ln_yesno(1)
/PATTERN t_groesse(1.3376) histgrad(3) ln_yesno(1)
/PATTERN t_groesse(1.3376) histgrad(4) ln_yesno(1)
/CRITERIA=PIN(.05) POUT(.10) ITERATE(20)
/PRINT=CI(95).

```

备注：根据 COXREG 子命令，给出生存时间变量（UEBERLEB）。与 SPSS 相比，利用 CATEGORICAL 子命令将变量 HISTGRAD 和 LN_YESNO 宣布为定类尺度。根据 METHOD 子命令，在 ENTER（方法“进入”）后面给出定量变量 T_GROESSE 以及两个分类变量 HISTGRAD 和 LN_YESNO。根据 METHOD 命令，只允许包括在上述模型（PATTERN，见下文）中考虑到的变量。利用 OUTFILE=命令，将“生存表”写入 SPSS 文件“cox_surv.sav”。关键词 TABLE 使得所创建的文件含有生存时间（变量“生存时间（月）”）、生存基线函数及其相应的平均值标准误差、分类变量（模式）的各个组合的生存（例如 HISTGRAD=2 和 LN_YESNO=0）及其相应的标准误差。同样，可以对累积危险函数输出基线、平均值函数以及不同的模式。

分类变量（模式）的各个组合通过 PATTERN 子命令传递给 SPSS。对于每个 PATTERN 子命令，SPSS 在图示中输出一条单独的直线。对变量 T_GROESSE、HISTGRAD 和 LN_YESNO 的效应进行分析并用图形表现出来，从而首先确定定量变量 T_GROESSE 的平均值（参见上文）。然后将分类变量的现有变量类别的数量相乘，对于 HISTGRAD 和 LN_YESNO 这两个变量得出的乘积就是 6。只要所有变量组合对于分析都是必要的（如果数据量很小，则不一定是这种情况），这就表明，必须向 SPSS 传递总共 6 个 PATTERN 子命令。

含有 PATTERN 子命令的第一行含有第一个模式，也就是说在定距尺度变量 T_GROESSE 后面含有其平均值、在第一个分类变量（HISTGRAD）后面含有其现有的第一个变量类别（在本例中：2），然后在第二个分类变量（LN_YESNO）后面含有其现有的第一个变量类别（在本例中：0）。对于下一个模式，根据同样的图进行流程。在定距尺度变量后面给出平均值；在分类变量后面给出现有的下一个变量类别，等等。重复这个过程，直到所有（相关的）变量组合（模式）都已经以含有 PATTERN 子命令的行的形式传递给 SPSS。

如果在 METHOD 后面给定了没有在模式中考虑到的变量，则 SPSS 仍将这些变量纳入分析，从而可能使对结果进行解释变得非常困难。在这种情况下，应将这些变量从 METHOD 后面的列表中删除，或者纳入含有 PATTERN 子命令的横行。如果没有利用 CATEGORIAL 子命令将相对于 SPSS 分类变量宣布为分类的，则这些变量将作为定量尺度变量被纳入分析；结果就造成对协变量平均值和模式数值的估计失真。子命令 PATTERN 不能用于分析时间相依协变量。


```
GET FILE= " C:\cox_surv.sav " .

variable labels
  SUR_1    " Ueberleben für HISTGRAD=2 LN_YESNO=0 "
  SUR_2    " Ueberleben für HISTGRAD=3 LN_YESNO=0 "
  SUR_3    " Ueberleben für HISTGRAD=4 LN_YESNO=0 "
  SUR_4    " Ueberleben für HISTGRAD=2 LN_YESNO=1 "
  SUR_5    " Ueberleben für HISTGRAD=3 LN_YESNO=1 "
  SUR_6    " Ueberleben für HISTGRAD=4 LN_YESNO=1 " .
exe.
```

根据表“分类变量的编码”和“协变量平均值和模式数值”分配变量标签。在分配变量 SURV_1,...,SURV_n 的标签时，应特别小心。在本例中，由于空间不足，放弃了在标签中对 T_GROESSE 的说明。因此，在解释结果时应注意，这些函数是基于三个变量，而不是像标签表面上看起来那样基于两个变量。

```
GRAPH
/SCATTERPLOT(OVERLAY)=UEBERLEB UEBERLEB UEBERLEB UEBERLEB
                        UEBERLEB UEBERLEB WITH SUR_1 SUR_2 SUR_3
                        SUR_4 SUR_5 SUR_6 (PAIR)

/MISSING=VARIABLE .

list var= UEBERLEB SUR_1 SUR_2 SUR_3 SUR_4 SUR_5 SUR_6 .
```

输出结果

平均值

报告

肿瘤大小		
平均值	标准误差	N
1.3376	1.27055	152

T_GROESSE 的平均值为 1.3376。对于各个模式的定义需要这个数值（见上文）。

Cox 回归

个案处理评估

		N	百分比
分析时的可用个案 未使用的个案	事件 ^a	66	38.6%
	已截尾总共缺失了数值	86	50.3%
	的个案	152	88.9%
	负数时间的个案	19	11.1%
	一个班次中出现最早事	0	0.0%
	件之前的截尾个案	0	0.0%
	总数	19	11.1%
		171	100.0%

a. 因变量：生存时间（月）

不再阐述“个案处理评估”表。

分类变量的编码^{b,c}

	频率	(1)	(2)
histgrad ^a 2=2	61	1	0
3=3	43	0	1
4=缺失数据	48	-1	-1
Ln_yesno ^a 0=No	116	1	
1=Yes	36	-1	

- a. 指示参数的编码
b. 分类变量: histgrad (组织学分级)
c. 分类变量: Ln_yesno (淋巴结)

对于表“协变量平均值和模式值”和变量标签分配的解释,表“分类变量的编码”是必不可少的。不再阐述表“分类变量的编码”。

组块 0: 初始组块

模型系数似然比检验

-2 对数似然值
572.873

不再阐述表“模型系数似然比检验”。

组块 1: 方法 = 进入

模型系数似然比检验^{a,b}

-2 对数 似然值	总 (值)			源自上一个步骤的变化			源自前一个组块的变化		
	卡方	df	显著性	卡方	df	显著性	卡方	df	显著性
499.512	103.130	4	0.000	73.361	4	0.000	73.361	4	0.000

- a. 初始组块编号 0, 先验对数似然值函数, -2 对数似然值: 572.873
b. 从组块 1 开始。方法 = 进入

不再阐述表“模型系数似然比检验”。

不再阐述“生存表 (没有显示)”。

方程中的变量

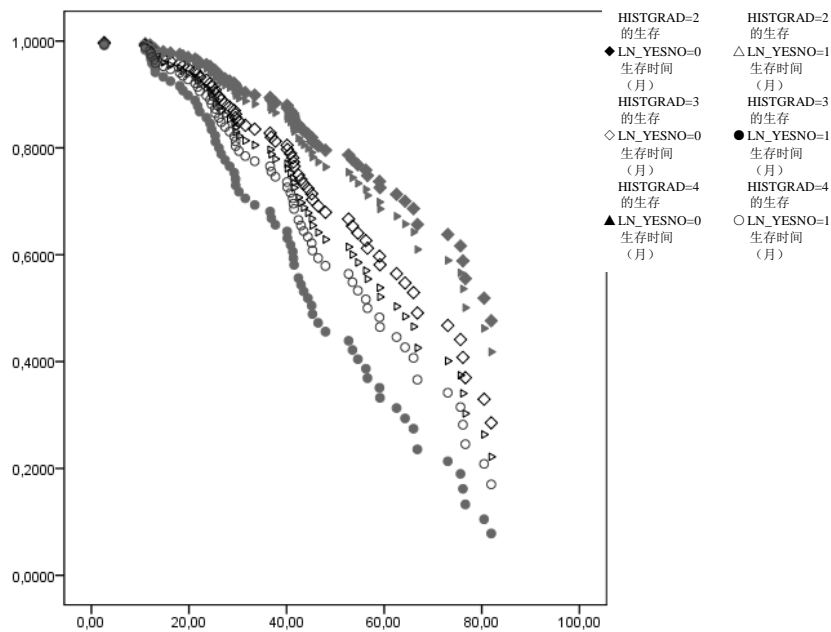
	B	SE	Wald	df	显著性	Exp(B)	Exp(B)的 95%置信区间	
							下限	上限
t_groesse	0.593	0.077	58.809	1	0.000	1.809	1.555	2.105
histgrad			3.405	2	0.182			
histgrad(1)	-0.229	0.179	1.632	1	0.201	0.795	0.559	1.130
histgrad(2)	0.296	0.174	2.911	1	0.088	1.345	0.957	1.890
ln_yesno	-0.354	0.134	6.947	1	0.008	0.702	0.539	0.913

分类变量 LN_YESNO 达到了统计显著性。置信区间剔除了 1。相反，分类变量 HISTGRAD 没有达到统计显著性。围绕着 Exp(B)的置信区间分别包括了 1。尽管输出的参数 T_GROESSE 和 LN_YESNO 看起来相等，但是由于其尺度不同，因此应对其进行不同的解释。不再阐述表“方程中的变量”。

协变量平均值和模式值

	平均值	模式					
		1	2	3	4	5	6 上限
t_groesse	1.338	1.338	1.338	1.338	1.338	1.338	1.338
histgrad(1)	0.086	1.000	0.000	-1.000	1.000	0.000	-1.000
histgrad(2)	-0.033	0.000	1.000	-1.000	0.000	1.000	-1.000
ln_yesno	0.526	1.000	1.000	1.000	-1.000	-1.000	-1.000

“协变量平均值和模式值”反映了 SPSS 内部使用的编码。如果回溯表格“分类变量的编码”（见上文），则可以发现，模式 1 相当于变量 TGROESSE=1.338、HISTGRAD=2 和 LN_YESNO=0 的组合。模式 2 相当于变量 TGROESSE=1.338、HISTGRAD=3 和 LN_YESNO=0 的组合。表格中的各个模式 1, ..., n 被 SPSS 依次输出为 SUV_1, ..., SUV_n。在分配 SUV 变量的标签时，应特别小心。不再阐述这个表格。



这个图反映了由模型预测的生存时间，以及针对各模式的“平均水平”患者所预测的风险。例如，对于具有 HISTGRAD=4 和 LN_YESNO=1 模式的“平均水平”患者而言，生存时间比具有 HISTGRAD=2 和 LN_YESNO = 0 模式的“平均水平”患者明显缩短得更快（分别考虑到了 T_GROESSE 的效应）。

ueberleb SUR_1 SUR_2 SUR_3 SUR_4 SUR_5 SUR_6

2.63	.9976	.9959	.9972	.9951	.9917	.9942
11.03	.9951	.9918	.9943	.9901	.9834	.9884
...省略了输出结果...						
76.63	.5554	.3699	.5008	.3028	.1326	.2454
80.47	.5187	.3295	.4621	.2636	.1049	.2085
81.93	.4764	.2854	.4181	.2217	.0783	.1701

Number of cases read: 64 Number of cases listed: 64

最后，输出具有相应模式的“平均水平”患者的生存时间估计量。

此外，利用 SPSS 可以给这些图加上置信区间。操作方法请参见关于 Kaplan-Meier 法生存分析的一章；该章有一个例子是针对累积生存概率测定 95% 置信区间。

其他形式

分别针对分类变量和定量变量的生存函数：

```
COXREG
  ueberleb
/STATUS=uestatus(1)
/CATEGORICAL = histgrad
/METHOD=ENTER t_groesse histgrad
/PATTERN t_groesse(1.3376) BY histgrad
/CRITERIA=PIN(.05) POUT(.10) ITERATE(20)
/PRINT=CI(95).
```

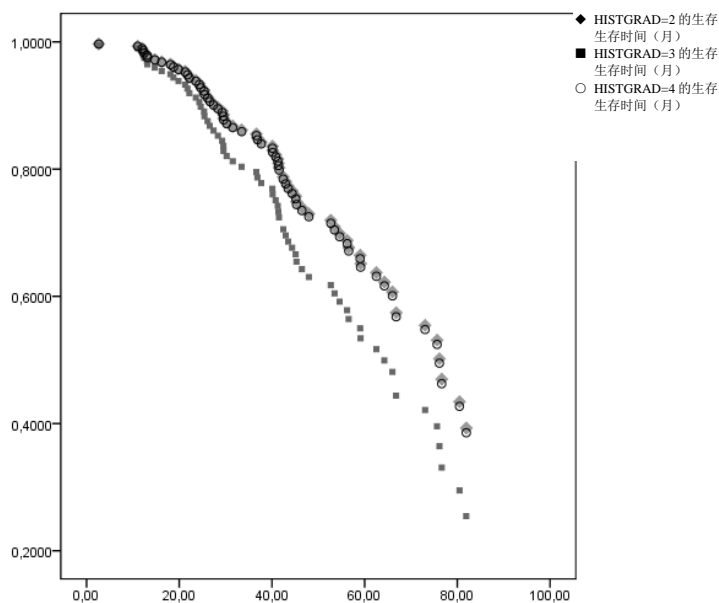
备注：如果只有一个分类变量，则可以使用一个 BY 命令，从而简化模式向 SPSS 的传递。在分析具有很多变量类别的分类变量时，尤其建议采用这种操作方式。但是，BY 命令需要通过/CATEGORICAL 子命令相对于 SPSS 将分类变量宣布为定类尺度。

```
GET FILE= " C:\cox_surv.sav " .

variable labels
  SUR_1      " Ueberleben für HISTGRAD=2 "
  SUR_2      " Ueberleben für HISTGRAD=3 "
  SUR_3      " Ueberleben für HISTGRAD=4 " .
exe.

GRAPH
/SCATTERPLOT(OVERLAY)= UEBERLEB UEBERLEB UEBERLEB
                     WITH SUR_1 SUR_2 SUR_3  (PAIR)
/MISSING=VARIABLE .
```

输出结果



备注：在图中，基本看不到 HISTGRAD=4 的生存函数，因为其差不多完全被 HISTGRAD=2 的生存函数覆盖。

4.7.6 检验 Cox 回归的前提条件

下面的例子演示了如何检验 Cox 回归的各个前提条件。这里归纳的各项标准主要是基于 Kleinbaum & Klein (2005) 和 Hosmer & Lemeshow (1999, 主要是第 6 章) 的著作。

概述：

- 基本数据
- 单变量和多变量离群值
- 截尾和目标事件 (Events) 之间的区别
- 多重共线性
- 对比例性假设的探索 (三个方法)
- 确定必要的样本范围 (示例)

基本数据

从表“个案处理评估”可以看出，可用于分析的个案的数量是否足够。

不应出现缺失值。但是如果出现缺失值，则应是完全随机分布的。

应根据具体环境的不同，检验对于时间相关方法典型的假设，即调查研究的具体环境保持不变，因此所有的边界因素在实验开始时、过程中和结束时都是可以相互比较的，但是根据以往的经验很难证明这个假设成立。

单变量和多变量离群值

离群值对推断性统计的估计量产生负面影响。因此，对于这些数据应检验是否有单变量和多变量离群值。形式上醒目的数值并不一定是内容上与众不同的数值（尤其是在社会科学领域

的数据中，关于离群值的定义和处理方法请参见 Schendera 的著作，2007）。

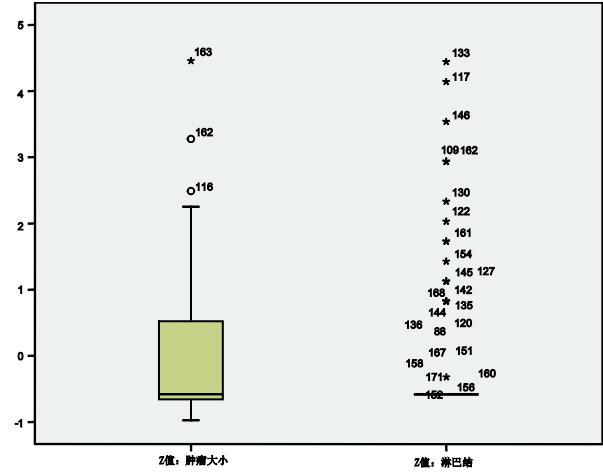
单变量离群值：

```
DESCRIPTIVES
  VARIABLES=t_groesse lym_knot
  /SAVE
  /STATISTICS=MEAN STDDEV MIN MAX.
DESCRIPTIVES
  VARIABLES=zt_groesse zlym_knot
  /STATISTICS= MIN MAX .
```

描述性统计量

	N	最小值	最大值
Z 值：肿瘤大小	152	-0.97404	4.45667
Z 值：淋巴结	171	-0.38266	5.64817
有效数值（成列）	152		

```
EXAMINE
  VARIABLES=Zt_groesse
  Zlym_knot
  /COMPARE VARIABLE
  /PLOT=BOXPLOT
  /STATISTICS=NONE
  /NOTOTAL/ID=id
  /MISSING=PAIRWISE .
```

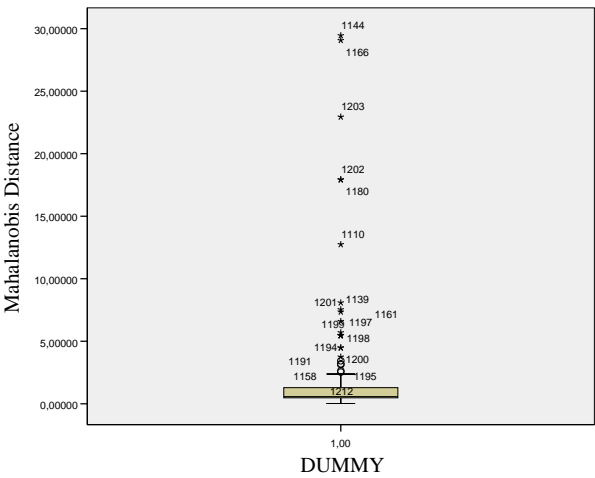


根据内容上的和/或形式上的标准不同，应利用超过某个大小（例如 3）的数值和/或其他特征更准确地对个案进行检验、转化或者剔除。

多变量离群值：

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT id
  /METHOD=ENTER t_groesse lym_knot
  /SAVE MAHAL .
compute DUMMY=1.
exe.
EXAMINE
```

```
VARIABLES=MAH_1 BY DUMMY /PLOT=BOXPLOT
/STATISTICS=NONE
/NOTOTAL/ID=id .
```



根据内容上的和/或形式上的标准不同，应利用某些特征更准确地对个案进行检验、转化或者剔除。

截尾和目标事件（Events）之间的区别

生存分析的假设是，具有截尾的一组个案是由于意外的原因对这项调查研究失效而被截尾的，与具有目标事件的一组个案没有系统性的区别。

```
COMPUTE event_zens=uestatus.
EXE.

REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT event_zens
/METHOD=ENTER
t_groesse lym_knot .
```

系数^a

模型		非标准化系数		标准化系数	T	显著性
		B	标准误差	Beta		
1	（常数）	0.014	0.035		0.398	0.691
	肿瘤大小	0.294	0.020	0.751	14.715	0.000
	淋巴结	0.025	0.009	0.141	2.764	0.006

a. 因变量：event_zens。

由于这个是多重检验，因此 α 系数（例如 0.05，预设置）除以需检验协变量的数量 n ($\alpha_{\text{kor}} = \alpha / n$)。输出的显著性明显低于 $\alpha_{\text{kor}} = 0.025$ 。可以得出的结论是，根据用回归分析方法检验的变量 T_GROESSE 和 LYM_KNOT，具有目标事件的个案与具有截尾的个案在统计学上有显著区别（也可以使用辨别分析）。可能的原因是系统性的，也就是非偶然的数据损失。如果这个解释正确，则甚至会改变最初真实的实验的状态：最初真实的实验在这次调查研究结束时不再真实，因为在实验结束时的数据不再是随机过程的产物，而是随着时间的推移由于一个未知偏误或者因素而失真。

原因可能是，随着这项调查研究的进行，具体环境发生了变化，从而使具有某些特征的个案由于意外的原因而对于这些调查研究失效了。可以猜测的是，引发这个现象的环境因素与个案的某些特征具有一种尚需人们探索的关联。

多重共线性

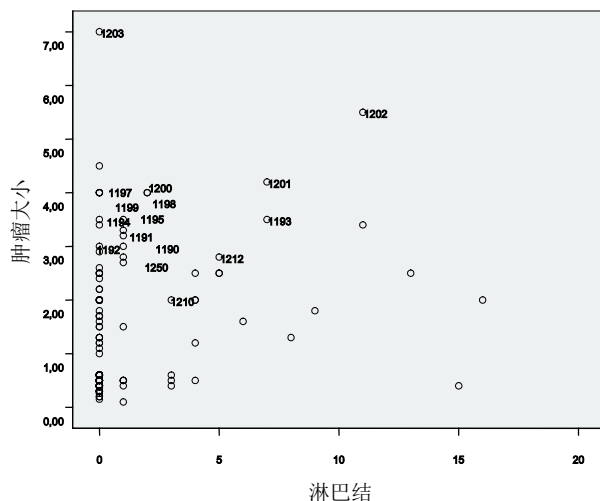
尤其是利用定量协变量的生存分析，特别容易受到内相关性或者多重共线性的影响。

GRAPH

```
/SCATTERPLOT(BIVAR)=lym_knot
WITH t_groesse BY id (IDENTIFY)
/MISSING=LISTWISE .
```

FACTOR

```
/VARIABLES t_groesse lym_knot
/MISSING LISTWISE
/ANALYSIS t_groesse lym_knot
/PRINT INITIAL EXTRACTION
/PLOT EIGEN
/CRITERIA MINEIGEN(1) ITERATE(100)
/EXTRACTION PAF
/ROTATION NOROTATE
/METHOD=CORRELATION .
```



例如，可以利用主轴因子分析检验多重共线性（见上文）。应先检验公因子方差 (SMC) >0.90 的变量，必要时将其从分析中剔除。例如，可以通过相关分析（SPSS 过程 CORRELATIONS）和补充性的通过散点图（见上文）检验简单的、成对的内相关性。相关性可能会因为影响力大的数值而产生巨大的失真。

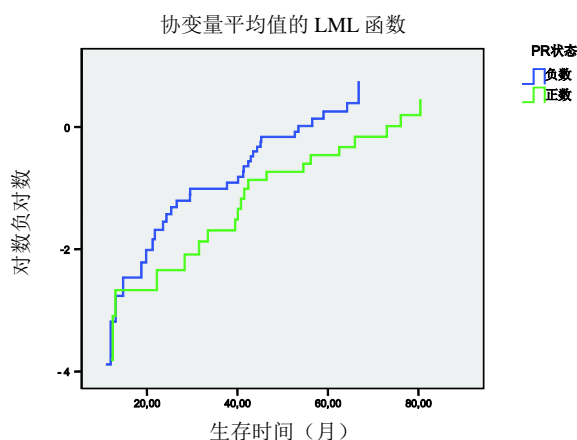
对比比例性假设的探索（三个方法）

Cox 回归认为，不同组患者生存时间的变化过程随着时间的推移是相同的，表现为基本平行的函数曲线。在 SPSS 中，可以首先利用 LML 图（LML: Log-Minus-Log，对数负对数）探索对比比例性假设的检验，然后利用明确的交互作用验证进行推断性统计检验（参见第 4.7.7 节）。下面给出的示例语句估计出一个分层模型。变量 PR_STAT 只被视为分层变量 (STRATA=)，但是不纳入模型 (METHOD=)。

如果风险具有比例性的假设成立，则 LML 图的各条直线应是平行的，绝不能相交。如果各条曲线相交，则必须否定比例性假设。对于定量协变量，建议探索性地分为比较小，但是有意义的类别（Klein & Moeschberger, 2003, 272-273）。如果有多个协变量，则可以检验各个类别的组合。

语句：

```
COXREG
  ueberleb /STATUS=uestatus(1)
  /METHOD=ENTER lym_knot
  /STRATA pr_stat
  /PLOT LML
  /PRINT=CI(95) BASELINE
  /CRITERIA=PIN(.05)
  POUT(.10) ITERATE(20) .
```



风险具有比例性的假设不成立，因为 LML 图的各条直线相交，所以必须否定比例性假设。

另一个备选的方法（参见 Kleinbaum & Klein, 2005, 546-548; Rasch 等人, 1998, 825-826）是基于 Kaplan-Meier 估计。Kaplan-Meier 法对生存时间（保存为 SUR_1）进行估计，将

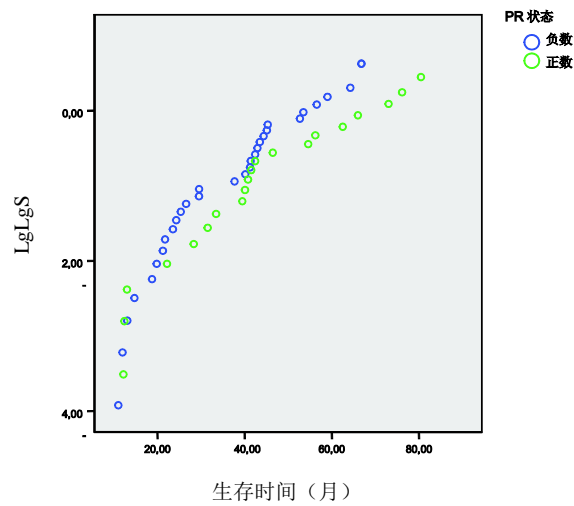
生存时间两次（一次负数）进行对数化，然后输出为一个分组的散点图。并且在这里，图中各条直线之间的距离应保持不变（平行）。

语句：

```
KM
  ueberleb BY pr_stat
  /STATUS=uestatus(1)
  /PRINT TABLE MEAN
  /SAVE=SURVIVAL .

compute LgLgS=ln(-ln(SUR_1)).
exe.

GRAPH
  /SCATTERPLOT(BIVAR)=
  ueberleb WITH LgLgS BY pr_stat
  /MISSING=LISTWISE .
```



风险具有比例性的假设不成立，因为图中的各条直线相交。

多变量 Cox 模型的另一个方法是通过 PRESID 测定出偏残差（又称 Schoenfeld 残差，参见 Kleinbaum & Klein, 2005, 150-153；也可参见 Hess, 2007）。在皮尔逊相关性中，最后的显著性检验了零假设，即没有违背风险具有比例性的假设。如果相关性检验得出了统计上的显著性，则违背了比例性假设。

语句：

```
COXREG
  ueberleb /STATUS=uestatus(1)
  /METHOD=ENTER t_groesse lym_knot
  /SAVE=PRESID
  /CRITERIA=PIN(.05) POUT(.10)
  ITERATE(20) .

RANK VAR=ueberleb (a)
  /rank
  /print=yes
  /ties=mean.

filter off.
use all.

CORRELATIONS
  /VARIABLES=
```

```
select if (uestatus=1).          rueberle pr1_1 pr2_1
exe.                             /PRINT=TWOTAIL NOSIG
                                /MISSING=PAIRWISE.
```

相关性

		生存的秩	t_groesse 的偏残差	lym_knot 的偏残差
生存的秩	皮尔逊相关性	1	0.329**	0.152
	显著性（两侧）		0.007	0.222
	N	72	66	66
t_groesse 的偏残差	皮尔逊相关性	0.329**	1	0.051
	显著性（两侧）	0.007		0.682
	N	66	66	66
lym_knot 的偏残差	皮尔逊相关性	0.152	0.051	1
	显著性（两侧）	0.222	0.682	
	N	66	66	66

变量 T_GROESSE 得出了统计上的显著性，在这里很明显违背了比例性假设。变量 LYM_KNOT 没有达到统计上的显著性，没有任何证据能够否定风险具有比例性的假设。

在进行解释时，应将样本大小也纳入观察范围。如果样本较大，则微小的误差也会变得显著；如果样本较小，即使较大的误差可能也不会变得显著。

确定必要的样本范围（示例）

下面这个确定必要的样本范围的公式适用于在一项随机调查研究中比较两个治疗组的 Cox 模型。下面的操作方法以 Rasch 等人的著作为基础（1998，816-817）。

提出的问题

在危险函数呈比例的假设下，针对两个相同大小的治疗组的比较和一个预设的风险比（HR=2），应首先确定两个治疗组预计的、有效的样本范围（“N_{effektiv}”），从而在给定的显著性水平（α=5）下这个风险比的检验优度大于或者等于 1-β（例如 100-20=80）。

但是，测定的样本范围（“N_{有效}”）没有考虑到截尾的出现。在第二步中，对于估计的 50%未截尾观测值（“N_{未截尾}”）测定整体的样本范围（“N_{整体}”）。

公式

$$N_{\text{effektiv}} = \frac{(u_{1-\alpha} - \sigma / 2 + u_{1-\beta} - \sigma / 2)^2}{(\log HR)^2 (1 - N_A) \times N_A}$$

其中：

- u_{1-α}=1-α/标准正态分布的第 2 象限(α =0.05)
- u_{1-β}=1-β/标准正态分布的第 2 象限(β=0.20)
- logHR=风险比的对数(HR)
- N_A=属于组 A 的患者所占比例

$$N_{\text{总共}} = N_{\text{有效}} \times p_{\text{未截尾}}$$

其中

- N_{总共}：总范围
- N_{有效}：有效的样本范围
- p_{未截尾}：未截尾观测值的概率

计算

$$N_{\text{有效}} = \frac{(1.96 + 0.84)^2}{(0.693)^2 \times (1 - 0.5) \times 0.5}$$

$$= \frac{7.84}{(0.693)^2 \times 0.25} = \frac{7.84}{0.12} \approx 65$$

$$N_{\text{总共}} = 65 \times (1/50 \times 100) = 130$$

其中:

$$u_{1-\alpha} = 1.96$$

$$u_{1-\beta} = 0.84$$

lnHR=2 的自然对数得出 0.693

$$N_A = 0.5 \text{ (相当于 50\%)}$$

其中

$$N_{\text{有效}}: 65$$

$$p_{\text{未截尾}}: 0.50 \text{ 或者 } 50\%$$

从这个计算中可以看出两方面的信息:

- 对于给定的参数, 需要的总样本范围为 $N=130$;
- 未截尾数据的比例越小, 所需的样本范围就越大。在未截尾个案只有 30% 时, 需要大约 $N=217$ 个个案 (参见 Rasch 等人著作, 1998)。

4.7.7 带有时间相依的定量协变量的 Cox 回归

Cox 模型 1 的基本假设有两个方面: a) 风险的比例性随着时间的变化保持不变。b) 一个个案的风险始终与任何一个其他个案的风险成比例 (也就是说, 比例常数与时间不相关)。可以将两组个案或者多组个案中的这个假设为二次对数生存率 (LML) 图中的平行线。根据研究领域 (市场调研、临床医学) 不同, 这个假设可能不是始终存在的。

例如, 在市场调研领域, 对于不同的促销措施就会有这样的情况。有些促销活动的效果在开始时非常明显, 但是随着时间的推移 (由于人们对此已经习惯) 效果越来越小。与之形成鲜明对比的是, 还有一些截然相反的促销活动, 其效果是逐渐累积的, 随着时间的推移效果越来越明显。如果将这两种促销活动所吸引的顾客人数进行比较, 则根据完全相反的趋势方向就可以认为, 比例性假设不成立。

而且在医学领域, 风险具有比例性的假设也不是始终恰当的。例如, 假设两组患者患有一种恶性肿瘤 (这个例子摘录自 Kleinbaum & Klein 的著作, 2005)。第一组患者接受了外科手术治疗, 这种措施尽管有一定风险, 但是被视为非常有效。第二组患者没有接受外科手术治疗, 而是采用了风险一般, 但是“只有”一般疗效的放射疗法。现在对于两组患者在开始治疗 (外科手术 vs. 化疗) 之后的风险变化过程可以如何想象? 第一组患者在开始手术后, 由于外科介入的风险高起初受到比较高的风险, 但是很快 (只要手术成功) 就转变为有效的痊愈。人们可以把这种风险想象成 S 形递减的阶梯函数。但是, 第二组患者由于放射, 受到的风险越来越大。为什么? 放射疗法的效果比外科手术小, 肿瘤的风险随着时间的推移越来越大。人们可以把这种风险想象成 S 形递加的阶梯函数。这两种危险函数的曲线走向完全相反, 甚至可以在任意一个点相交。这个例子就表明, 风险具有比例性的假设是不成立或者不恰当的。

利用 SPSS 可以检验风险具有比例性的假设 (图形式探索方法参见第 4.7.6 节), 并用两种方式进行校正 (例如参见 Kleinbaum & Klein, 2005, 153-157): 对于每个协变量, 建立与

一个时间相依变量的交互作用，然后将其纳入模型。如果这些交互作用不显著，则无法否定风险具有比例性的假设。相反，如果这些交互作用是显著的，则只能在连同交互作用项一起的情况下进行最初的模型检验。另外一种方法是，可以用一个定量协变量[只要其测量水平（即定距尺度）的信息不是这次调查研究的核心内容]及其与时间相依协变量的交互作用建立一个分层变量，然后将其纳入模型。下面的例子展示了如何利用 Cox 回归检验风险具有比例性的假设，以及如何将 Cox 回归检验用于可能的时间相依协变量（即 Cox 模型 2）。

方法 1：一个时间相依的定量协变量

提出的问题

要对定量变量 T_GROESSE（肿瘤尺寸）对于患者生存时间的影响进行建模。假设是，肿瘤尺寸可能是时间相依的（随着时间的推移，肿瘤越来越大），因此可能违背了这个模型的风险具有比例性的假设。

图示主要集中在当前分析实例的关键部分。已知的界面选择路径、SPSS 语句中记录的设置、完全相同的或者至少十分类似的输出结果在这里不再阐述或者解释。

COXREG 命令利用一个系统内置的时间变量（“T”），可以检验随着时间的变化是否有影响，也就是说，比例性风险的模型所基于的假设是不成立的，可能无法应用这个模型。接下来的处理分为三个步骤。

- 在第一次 Cox 回归（没有之后测得的时间相依协变量“T_COV_”）时测定偏残差，并将其另存为 PR_1。
- 在一个双变量散点图中，在 X 轴上截取生存时间，在 Y 轴上截取 T_GROESSE 的偏残差。这个散点图作为图形式的前提条件检验，不应具有任何特殊之处。
- 在第二次 Cox 回归时，在 T_GROESSE 后面测定协变量“T_COV_”，并将其纳入模型。

语句：

```
COXREG
  ueberleb /STATUS=uestatus(1)
  /METHOD=ENTER t_groesse
  /SAVE=PRESID
  /PRINT=CI(95)
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) .

GRAPH
  /SCATTERPLOT(BIVAR)=ueberleb WITH PR1_1
  /MISSING=LISTWISE .

TIME PROGRAM.
COMPUTE T_COV_ = T_*t_groesse .
COXREG
  ueberleb /STATUS=uestatus(1)
  /METHOD=ENTER t_groesse T_COV_
  /PRINT=CI(95)
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) .
```

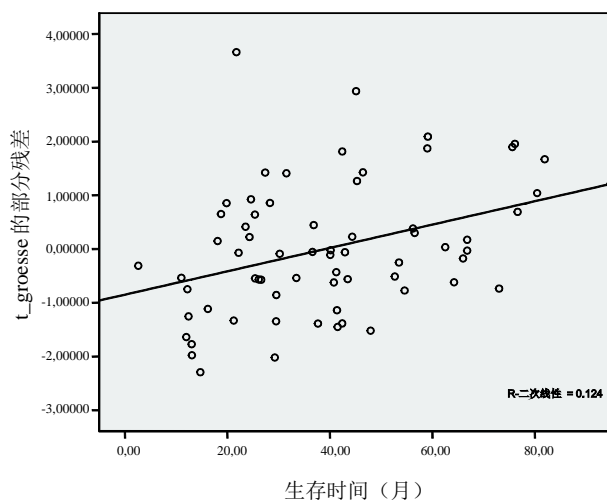
备注：通过 `SAVE=RESID` 在第一次 Cox 回归时测定偏残差，然后将其保存在变量 `PR1_1` 中。在双变量散点图（`SCATTERPLOT`）中，在 X 轴上截取生存时间 `UEBERLEB`，在 Y 轴上截取偏残差（`PR1_1`）。通过 `TIME PROGRAM` 对系统内置的时间变量“`T_`”进行定义。所创建的、同样是时间相依的协变量 `T_COV_` 就是基于 `T_` 和 `t_groesse` 的乘积（参见 `COMPUTE`）。在紧接着的第二次 Cox 回归时，根据 `METHOD` 命令在 `T_GROESSE` 后面将协变量 `T_COV_` 纳入模型。时间相依协变量的 Cox 回归不能输出图形。

Cox 回归——包括 `t_groesse`

这里没有反映出完整的输出结果（例如对个案处理的评估、方程中的变量等）。这个计算只能用于存储偏残差，并将其用散点图的形式展现出来。

散点图已经做了编辑。回归直线是后来手工添加的。

图



根据散点图，检验风险具有比例性的假设。在 X 轴上截取生存时间，在 Y 轴上截取表示肿瘤尺寸的偏残差。

一个协变量的偏残差是每个个案的协变量观测值和期望值（根据指定模型；前提条件是这个模型是正确的）之间的差别。只针对未截尾个案测定偏残差，因此，散点图中的点只基于未截尾个案。

如果关于肿瘤尺寸的风险具有比例性的假设是正确的，则这个散点图不应含有任何显著的分布。这个图只表明偏相关和时间之间具有微弱的正相关，从而让人猜测，肿瘤的尺寸看起来只在很小程度上受到时间的影响。为了检验这种猜测，给模型添加一个时间相依协变量。

Cox 回归——`t_groesse` 和 `T_COV_`

不再阐述表“个案处理评估”。

不再阐述“组块 0”的表“模型系数综合检验”。

组块 1：方法 = 进入

模型系数综合检验

-2 对数似然值	总值			由于先前步骤产生的变化			由于先前组块产生的变化		
	卡方	df	显著性	卡方	df	显著性	卡方	df	显著性
503.259	106.304	2	0.000	69.614	2	0.000	69.614	2	0.000

不再阐述表“模型系数综合检验”。

表“方程中的变量”表明，所创建的时间相依变量 T_COV_ 达到了统计显著性（ $p=0.030$ ）。相应的系数 B 尽管很小（0.009）。但是 B 没有编辑者，可能有很大的误导作用。此外， $\text{Exp}(B)=1$ 没有进入 95% 置信区间。据此，协变量 T_GROESSE 对 UEBERLEB 的效应与时间具有统计学意义上的相关性。必须否定风险具有比例性的假设。只能在连同交互作用项一起的情况下进行最初的模型检验。另外一种方法是，可以用一个定量协变量及其与时间相依协变量的交互作用建立一个分层变量，然后将其纳入模型。

方程中的变量

	B	SE	Wald	df	显著性	Exp(B)	Exp(B) 的 95% 置信区间	
							下限	上限
t_groesse	0.327	0.162	4.060	1	0.044	1.386	1.009	1.905
T_COV_	0.009	0.004	4.728	1	0.030	1.009	1.001	1.017

方法 2：多个时间相依的定量协变量

如果要对多个时间相依的定量协变量的影响进行建模，则可以通过时间变量 T_ 的对数检验风险具有比例性的假设。

提出的问题

要对定量变量 T_GROESSE（肿瘤尺寸）和 LYM_KNOT（淋巴结）对于患者生存时间的影响进行建模。这两个变量在一定情况下是时间相依的，可能违背了这个模型的风险具有比例性的假设。

语句：

```
COXREG
  ueberleb /STATUS=uestatus(1)
  /METHOD=ENTER t_groesse lym_knot
  /SAVE=PRESED
  /PRINT=CI(95)
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) .

GRAPH
  /SCATTERPLOT(BIVAR)=ueberleb WITH PR1_1
  /MISSING=LISTWISE .

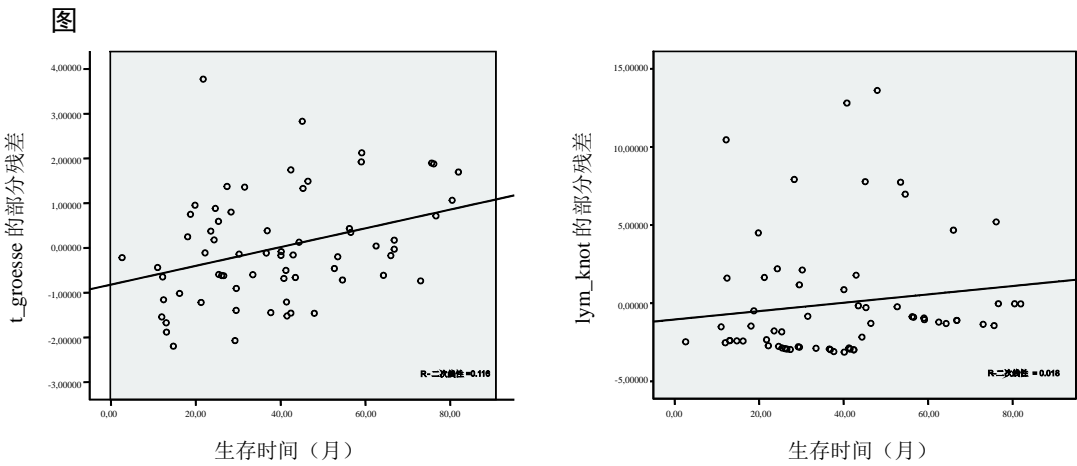
GRAPH
  /SCATTERPLOT(BIVAR)=ueberleb WITH PR2_1
  /MISSING=LISTWISE .
```

```
TIME PROGRAM.
COMPUTE T_COV_ = ln(T_) .
COXREG
  ueberleb /STATUS=uestatus(1)
  /METHOD=ENTER t_groesse
  t_groesse*T_COV_ lym_knot lym_knot*T_COV_
  /PRINT=CI(95)
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) .
```

备注：通过 COMPUTE 命令计算出系统内置时间变量“T_”的自然对数，将其创建为协变量 T_COV_。在第二次 Cox 回归时，在 METHOD 命令后面还要将 T_COV_与定量协变量 T_GROESSE、LYM_KNOT 之间的相互作用纳入模型。

Cox 回归——包括 t_groesse 和 lym_knot

这里没有反映出完整的输出结果（例如，对个案处理的评估、方程中的变量等）。这个计算只能用于存储偏残差，并将其用散点图的形式展现出来。



根据多个散点图，检验风险具有比例性的假设。这些图表明偏相关和时间之间只具有一定的正相关，从而让人猜测，肿瘤的尺寸看起来更像是与时间不相依。为了检验这种猜测，给模型添加了与时间相依协变量的交互作用。

Cox 回归——t_groesse、lym_knot 和 T_COV_

不再阐述表“个案处理评估”。

不再阐述“组块 0”的表“模型系数综合检验”。

组块 1：方法 = 进入

模型系数综合检验									
-2 对数 似然值	总值			由于先前步骤产生的变化			由于先前组块产生的变化		
	卡方	df	显著性	卡方	df	显著性	卡方	df	显著性
502.811	109.368	4	0.000	70.062	4	0.000	70.062	4	0.000

不再阐述表“模型系数综合检验”。

方程中的变量

	B	SE	Wald	df	显著性	Exp(B)	Exp(B)的 95%置信区间	
							下限	上限
t_groesse	-0.026	0.452	0.003	1	0.955	0.975	0.402	2.364
lym_knot	-0.157	0.257	0.373	1	0.541	0.855	0.517	1.414
t_groesse*T_COV_	0.185	0.132	1.977	1	0.160	1.204	0.930	1.559
lym_knot*T_COV_	0.057	0.073	0.601	1	0.438	1.058	0.917	1.222

输出的显著性明显超过 $\alpha_{\text{kor}}=0.05$ 。Exp(B)=1 进入 95%置信区间。由于这个是多重检验，因此 α 系数（例如 0.05，预设置）还要除以需检验交互作用的数量 n ($\alpha_{\text{kor}} = \alpha/n$)。从这个原因可以得出结论，任何一个协变量都与时间没有统计学意义上的相关性。风险具有比例性的假设无法否定。

4.7.8 Cox 回归的特定前提条件

1. 风险具有比例性。时间不相依协变量的 Cox 回归是基于这个假设：两个个案的风险比随着时间的推移保持不变。应始终检验协变量与时间不相依这个核心假设是否成立。如果风险没有比例性，则模型（相对风险、标准误差、Power）的预测量以及据此推导出的说法的可靠性都大幅降低。可以用不同的方式检验风险的比例性。最常用的是图形法，如 Schoenfeld 图。在这里，分别从 Y 轴上截取各自协变量的残差，从 X 轴上截取生存时间。对于两个或者多个患者组，可以调用风险图、LML（对数负对数）图或者生存函数图。对于不同的患者组，随着时间的推移风险具有比例性应表现为几条几乎平行的直线。如果这些直线相交，则违背了风险具有比例性的假设。LML 图不适合只有连续协变量的模型。也可以用数值法，即回归或者 Harrell's Rho 统计法。

如果风险没有比例性，则可以用不同的方式处理这种情况。例如，可以将分层变量纳入模型。也可以将生存时间分解为部分值域，对于这些值域分别估计其风险。例如，也可以将协变量和生存时间之间的交互作用项纳入模型，进行检验。对于每个协变量，建立与一个时间相依变量的交互作用，然后将其纳入模型。如果这些交互作用不显著，则无法否定风险具有比例性的假设。相反，如果这些交互作用是显著的，则只能在连同交互作用项一起的情况下进行最初的模型检验。此时，可以或者不可以转化生存时间，对于选择某个特定的转化函数没有具有约束力的规定，相应的，不同的函数对于风险比例性的假设完全有可能得出不同的结论（参见 Kleinbaum & Klein, 2005, 153-157）。另外一种方法是，可以用一个定量协变量（只要其不是这次调查研究的核心内容）及其与时间相依协变量的交互作用建立一个分层变量，然后将其纳入模型。

2. 随机样本。对基线危险函数的估计是根据数据推导出来的，因此应具有随机样本。如果没有随机样本（non-random samples），则这种现象会将巨大的偏倚带入参数预测量，从而导致危险函数和生存函数不准确（Boehmke, Morey & Shannon, 2006）。对于随机样本适用的规则是：越大，越好。对于小的样本，不应使用 Cox 回归，主要是防止因为过度拟合而产生人为臆造的危险（Box-Steffensmeier & Jones, 2004, 89; Box-Steffensmeier & Jones, 1997, 1434）。

3. 目标事件。Cox 回归的标准模型研究的是由一次性出现的事件造成的风险。这个模型不能与一次性出现多个等效（“竞争性”）事件（竞争风险生存分析）或者重复出现的事件（复发事件生存分析）的模型相混淆。偏似然法是建立在目标事件序列的基础上。因此，只有在个别情况下目标事件才能同时出现（被称为“结点”）结点的比例不应超过 5%。当结点很少时，Efron 方法比前面介绍的 Breslow 方法更加精确。
4. 模型的叠加。如果因变量根据自变量的数值发生了达到自变量一个单位的变化，则出现非线性。与此相比，如果因变量根据其他自变量的其中一个的数值发生了达到自变量一个单位的变化，则呈现非叠加性。例如，可以通过检验是否存在可信的或者理论上可能有的所有交互作用，就可以检验模型的叠加性。后一种方法只适用于相对简单的模型。
5. 分类变量的类别等级完整性。在有数据缺失的情况下，结果部分中在括号里的变量类别与编码不一致。例如，如果数据被编码为 0 至 7，但是只有变量类别 0、1、5 和 7 存在，则不显示编码（0）、（1）和（5），而是显示（1）、（2）和（3）（最高的编码是冗余的）。因此，建议仔细检验类别等级是否完整，以确定所测定的参数是否也与正确的类别等级相联系。例如，在这个例子中，给出的变量类别（1，第一级）可能与编码（1，第二级）相混淆。为了辨别清楚，应不断查看随之输出的“分类变量的编码”表。
6. 分类变量的参考类别。参考类别对所测定结果的高度和方向具有决定性影响。例如，风险比编码为 1 时可能数值达到 3，但是编码为 0 时可能达到 0.33。例如，对于二元因变量（B）的系数，正负号发生改变。SPSS 过程命令 COXREG 根据预设置，始终将分类变量的最后一种变量类别选择为参考类别。其他作者、分析师或者软件在有些情况下使用了其他的参考类别。请检查（自动）选择的参考类别是否符合评估目的，否则就将分类变量重新编码。

在临床或者流行病学研究中适用的规则是，始终将病例组（暴露，事件）编码为“1”，始终将对照组（不暴露，事件不出现）编码为“0”。
7. 定量协变量。在进行 Cox 回归时，定量下相互之间应不相关（排除多重共线性）。协变量之间的任何相关（例如 >0.80 ）都是多重共线性的迹象。通过参数估计量明显很高的标准误差（非标准化： >2 ，标准化： >1 ）显示出多重共线性。可以通过相关性或者主轴因子分析来检验内相关性或者多重共线性（参见第 4.7.6 节）。应先检验公因子方差（SMC） >0.90 的变量，必要时将其从分析中剔除。是否可以和在多大程度上可以消除多重共线性，除了相关协变量的数量和相关性之外，主要取决于错误出现在研究过程的哪个地方：理论构建、具体实施或者数据搜集。“如果发现了多重共线性，具体怎么处理更像是一门艺术，而不是科学”（Menard, 2001, 80；也可参阅 Pedhazur, 1982², 247）。
8. 模型设定。模型设定应是通过关于内容的统计学标准，而不是通过关于形式的算法来推导的。各个协变量应相互不相关。很多作者明确建议不要使用自动变量选择的方法，但是在有保留的情况下，作为一种探索性方法也是可以使用的。对于这两种不同的工作方法，建议采用下面的处理方式：首先接受一个内容上相关的协变量，通过显著性检验（T 统计量）剔除统计上无关的变量。如果模型含有协变量之间显著的交互作用，则影响达不到显著性的协变量也保留在模型中。

9. 例如, 逐步法是根据形式上的标准(统计学上的关联)进行工作的, 不适用于理论推导的建模, 因为逐步法也选择了内容上没有关联的协变量。应根据可信的、关于内容的标准, 对纯粹探索性的或者预测性的工作方法进行交互检验。后退法应优先于前进法, 因为后退法与前进法相反, 是从检验一阶交互作用开始的, 因此不存在仓促剔除潜在的抑制变量的风险。但是, 逐步法不消除多重共线性, 因此至少要通过交叉验证予以保障。
10. 在解释回归系数和风险比(Exp(B))时的特殊性。(a) 协变量的尺度水平: 对于风险比和回归系数的解释, 其区别在于分类预测量和定量预测量。对于定量变量, 可以用整个统一定义域的一个公共值来表达其影响; 对于分类变量, 则测定 $n-1$ 个变量类别或单位的数值。需要注意的是, 编码会对风险比或回归系数的大小或正负号起作用(例如, 二元因变量的系数可能正负号颠倒, 对此参阅关于参考类别的注释)。(b) 分类协变量的编码: 协变量的编码对回归系数的解释及其计算有影响。如果对于病例个案或者事件个案, 编码偏离 1, 并且对于对照个案, 编码偏离 0(对此参阅关于参考类别的备注), 则必须用另外的方法测定参数。(c) 非标准化回归系数对比标准化回归系数: 非标准化回归系数可能与标准化回归系数有很大的区别, 并且完全错误地反映了各自协变量的影响。Menard(2001)建议对于分类变量和带有自然单位的变量采用非标准化回归系数或风险比, 对于没有共同单位的定量尺度采用标准化回归系数。通过在分析之前将定量协变量本身标准化, 就可以针对定量协变量的模型提取出标准化回归系数。然后就可以将提取出的回归系数解释为标准化回归系数。
在线性回归中, 通常建议将标准化回归系数用于比较在一个样本/总体内部的定量变量, 或者用于没有共同单位的定量变量, 对于后者应考虑到, 其测定可能是取决于所选择的样本, 并且根据模型拟合优度不同, 只能有所保留地将这种测定结果普遍化。非标准化回归系数建议用于比较样本/总体之间的定量变量, 或者用于具有自然/共同单位的定量变量。根据这两种回归系数的优点和缺点, Pedhazur(1982², 247-251)建议给定两种度量。如果在分析之前将数据 z 标准化, 则将 β 值给定为 B 值。
11. 离群值。离群值可能对推断性统计的估计量产生负面影响。因此, 对于这些数据应检验是否有单变量和多变量离群值(关于离群值的定义和处理方法参见 Schendera, 2007)。
12. 残差分析。残差分析适用于评估 Cox 回归模型的优度和适用性(例如 Kleinbaum & Klein, 2005, 151-153; Klein & Moeschberger, 2003, 353-392; Hosmer & Lemeshow, 1999, 197-225)。SPSS 可以对残差分析保存三种不同的残差类型: Cox-Snell 残差、偏残差或者回归系数差别(只用于具有至少一个协变量的模型)。
 - 通过 Cox-Snell 残差可以识别出影响巨大的个案(SPSS SAVE 选项: HAZARD=RESID)。
 - 偏残差(也就是 Schoenfeld 残差)可以用推论统计学方法或者图形法对风险具有比例性的假设进行检验(参见第 4.7.6 节)。通过偏残差可以识别出影响巨大的个案(SPSS SAVE 选项: PRESID, 参见第 4.7.7 节)。
 - 如果删除了相应的个案, 则名为“回归系数差别”的残差反映了系数变化的幅度

（只用于具有至少一个协变量的模型（SPSS SAVE 选项：DFBETA）。

- 方便地自行生成的鞅残差可以对协变量的对数线性进行图形法的检验。如果散点图中的分布接近于线性，则协变量和危险函数的对数之间存在必要的对数线性。

可以通过 Cox-Snell 残差（又称累积危险函数，保存在变量 HAZ_1 中）和 STATUS 变量推导出鞅残差。假设 STATUS 变量被称为 EREIGNIS（事件），数值 1 定义了一个事件的出现，则可以通过下列 COMPUTE 步骤测定鞅残差，并保存在变量 MARTGALE 中。

```
compute MARTGALE= (EREIGNIS=1) - HAZ_1.  
exe.
```

据此，鞅残差和“相反的”Cox-Snell 残差没有什么区别。

通过散点图或者箱线图展示这些参数，就可以方便地识别出离群值。例如，在第 4.7.7 节展示了，如果根据偏残差可以检验风险具有比例性的假设（关于其他用途可以参见 Klein & Moeschberger, 2003, 第 11 章）。

4.7.9 附录：对比方法

SPSS 为 Cox 回归提供了下列对比方法（按字母顺序排列，括号中是相应的 SPSS 选项）：“偏差”（预设）、“简单”、“差别”、“Helmert”、“指示”、“多项式”、“特别”（只有通过语句才可用）和“重复”。在“偏差”和“简单”对比时，可以将第一个或者最后一个类别确定为参考类别。Helmert、差别和多项式对比是正交的，“特别”类型的对比是不必要的。

正交对比在统计学上不相依，没有冗余。如果（a）每一行的对比系数总和为 0，（b）在不相交的各行所有成对数值相应的系数乘积的总和同样为 0，则对比是正交的。这些阐述基本上是根据 SPSS V16 的技术资料得来的。

“偏差”（DEVIATION，预设置）

在采用“偏差”对比方法时，除了一个平均值超过所有因子等级之外，对每个因子等级进行比较。偏差对比具有下列形式：

平均值	(1/k 1/k)
df(1)	(1-1/k -1/k ... -1/k -1/k)
df(2)	(-1/k 1-1/k ... -1/k -1/k)
...	
...	
df(k-1)	(-1/k -1/k ... 1-1/k -1/k)

k 相当于自变量类别的数量。在标准设置中，删除最后一个类别。例如，具有三个类别的自变量的偏差对比如下：

(1/3 1/3 1/3)

$$\begin{pmatrix} 2/3 & -1/3 & -1/3 \end{pmatrix}$$

$$\begin{pmatrix} -1/3 & 2/3 & -1/3 \end{pmatrix}$$

如果不是删除最后一个类别，而是另一个类别，则在关键词 **DEVIATION** 后面的括号中给出需要删除类别的编号。例如，在下面的例子中，计算了第一个和第三个类别的偏差，删除了第二个类别。

/CONTRAST(FAKTOR)=DEVIATION(2)

如果因子有三个类别，则计算对比矩阵如下（区别在于第二个类别的编码）：

$$\begin{pmatrix} 1/3 & 1/3 & 1/3 \end{pmatrix}$$

$$\begin{pmatrix} 2/3 & -1/3 & -1/3 \end{pmatrix}$$

$$\begin{pmatrix} -1/3 & -1/3 & 2/3 \end{pmatrix}$$

“简单” (SIMPLE)

在采用“简单”对比方法时，将每个因子等级与最后一个因子等级进行比较。将一个因子的每一个等级（除了最后一个等级之外）与最后一个因子等级进行比较。一般的矩阵形式是：

平均值($1/k \ 1/k \ \dots \ 1/k \ 1/k$)

$$\text{df}(1) \quad \begin{pmatrix} 1 & 0 & \dots & 0 & -1 \end{pmatrix}$$

$$\text{df}(2) \quad \begin{pmatrix} 0 & 1 & \dots & 0 & -1 \end{pmatrix}$$

...

...

$$\text{df}(k-1) \quad \begin{pmatrix} 0 & 0 & \dots & 1 & -1 \end{pmatrix}$$

k 相当于自变量类别的数量。例如，具有四个类别的一个自变量的简单对比如下：

$$\begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & 0 & -1 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 1 & 0 & -1 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 0 & 1 & -1 \end{pmatrix}$$

如果不是使用最后一个，而是不同于参考类别的另一个类别，则在关键词 **SIMPLE** 后面的括号中给出当前参考类别的编号。这个编号不一定与这个类别的数值一致。例如，用子命令 **CONTRAST** 计算出一个对比矩阵，其中删除了第二个类别：

/CONTRAST (FAKTOR) = SIMPLE (2)

如果因子具有四个类别，则计算出下列对比矩阵：

$$\begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}$$

$$\begin{pmatrix} 1 & -1 & 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & -1 & 1 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & -1 & 0 & 1 \end{pmatrix}$$

“差别” (DIFFERENCE)

在采用差别对比方法（也就是相反的 Helmert 对比）时，将一个自变量的类别与变量先前类别的平均值进行比较。将一个因子的每一个等级（除了第一个等级之外）与先前因子等级的平均值进行比较。一般的矩阵形式是：

平均值 $(1/k \ 1/k \ 1/k \ \dots 1/k)$
 $df(1) \ (-1 \ 1 \ 0 \ \dots 0)$
 $df(2) \ (-1/2 \ -1/2 \ 1 \ \dots 0)$
 \dots
 \dots
 $df(k-1) \ (-1/(k-1) \ -1/(k-1) \ -1/(k-1) \ \dots 1)$

k 相当于自变量类别的数量。

例如，具有四个类别的一个自变量的差别对比如下：

$(1/4 \ 1/4 \ 1/4 \ 1/4)$
 $(-1 \ 1 \ 0 \ 0)$
 $(-1/2 \ -1/2 \ 1 \ 0)$
 $(-1/3 \ -1/3 \ -1/3 \ 1)$

“Helmert” (HELMERT)

在采用 Helmert 对比方法时，将一个自变量的类别与后续类别的平均值进行比较。将一个因子的每一个等级（除第一个等级之外）与后续因子等级的平均值进行比较。一般的矩阵形式是：

平均值 $(1/k \ 1/k \ \dots 1/k \ 1/k)$
 $df(1) \ (1 \ -1/(k-1) \ \dots -1/(k-1) \ -1/(k-1))$
 $df(2) \ (0 \ 1 \ \dots -1/(k-2) \ -1/(k-2))$
 \dots
 \dots
 $df(k-2) \ (0 \ 0 \ 1 \ -1/2 \ -1/2)$
 $df(k-1) \ (0 \ 0 \ \dots 1 \ -1)$

k 相当于自变量类别的数量。例如，有四个类别的一个自变量具有下列形式的对比矩阵：

$(1/4 \ 1/4 \ 1/4 \ 1/4)$
 $(1 \ -1/3 \ -1/3 \ -1/3)$
 $(0 \ 1 \ -1/2 \ -1/2)$
 $(0 \ 0 \ 1 \ -1)$

“指示” (INDICATOR)

指示对比也称为哑变量编码，在这里对比指的是是否从属于某个类别。重新编码的变量的数值等于 $k-1$ 。因此，一个二元变量只会得出一个变量；例如，具有 k 个变量类别的多等级变量会得出 $k-1$ 个变量。对于所有 $k-1$ 个变量，参考类别中的个案编码为 0。在 i -ten 类别中的一

个个案对于几乎所有的指示变量编码为 0，只对 i-ten 变量编码为 1。在对比矩阵中，参考类别显示为带有零的横行。

“多项式” (POLYNOMIAL)

在对趋势进行检验以及调出作用面积时，多项式对比特别有用。例如，多项式对比检验了是否在所有因子水平都具有线性或者二次关联。多项式对比也可以用于非线性曲线估计，例如用于曲线回归。在平衡的方案设计中，多项式对比是正交的。在正交的多项式对比中，第一个自由度含有所有类别的线性效应，第二个自由度含有二次效应，第三个自由度含有三次效应，依次类推。

这就可以认为，因子的各个等级之间的距离是相等的。但是对于“多项式对比”，也可以探索性地给定各个等级之间的距离。可以用从 1 到 k 的连续整数给定相同的距离（预设置），其中 k 等于类别的数量。如果变量 DOSIS 具有三个类别，则子命令 /CONTRAST (DOSIS) = POLYNOMIAL 相当于命令 /CONTRAST (DOSIS) = POLYNOMIAL (1, 2, 3)。但是并不总是存在相等的距离。假设 DOSIS 是一种三个不同组患者服用的有效物质的不同剂量。但是，如果第二组患者的服用剂量是第一组患者的四倍，第三组患者的服用剂量是第一组患者的七倍，则可得出下列编码：/CONTRAST (DOSIS) = POLYNOMIAL (1, 4, 7)。对于多项式编码，只有因子的各个等级之间的相对差异是重要的。POLYNOMIAL (1, 2, 4) 和 POLYNOMIAL (2, 3, 5) 或者 (20, 30, 50) 相同，因为在每个多项式对比中，第 2 个和第 3 个变量类别之间差别与第 1 个和第 2 个变量类别之间差别的比例是相同的。第 2 个和第 3 个变量类别之间差别是第 1 个和第 2 个变量类别之间差别的两倍。

“特别” (SPECIAL)

“特别”是可以由用户定义的对比方法。在这里可以用二次矩阵的形式给定一个特别对比，其中行和列的数量必须相当于自变量中的类别数量。

第一行通常构成平均值效应（恒定效应），是加权值的组合，从而说明了通过现有的变量如何测定其他自变量（只要有的话）的平均值。一般而言，这个对比是由多个 1 组成的向量。

矩阵的其他各行含有特别对比，其说明了在变量的各个类别之间的比较。通常，正交对比是最为有用的。正交对比在统计学上不相依，没有冗余。

示例：

一个因子具有 4 个等级，要对不同的等级进行相互比较。例如，对此就适用下列特别的对比方法：

(1 1 1 1) 计算平均值时的加权
 (3 -1 -1 -1) 1 与 2 至 4 的比较
 (0 2 -1 -1) 2 与 3 和 4 的比较
 (0 0 1 -1) 3 与 4 的比较

特别的对比方式应展现出各个等级非线性的相互组合。但是如果不是这样，则程序报告线性相依，取消数据处理。

“重复” (REPEATED)

“重复”反映了一个自变量前后相连等级的比较。将一个因子的每一个等级（除最后一个等级之外）与下一个因子等级进行比较。在进行剖面分析以及在需要差值的情况下，“重复”对比是非常有用的。一般的矩阵形式是：

平均值 $(1/k \ 1/k \ 1/k \ \dots 1/k \ 1/k)$

$df(1) \ (\ 1 \ -1 \ 0 \ \dots 0 \ 0)$

$df(2) \ (\ 0 \ 1 \ -1 \ \dots 0 \ 0)$

$\cdot \cdot \cdot$

$\cdot \cdot \cdot$

$df(k-1) \ (\ 0 \ 0 \ 0 \ \dots 1 \ -1)$

k 相当于自变量类别的数量。例如，具有四个类别的一个自变量的重复对比如下：

$(\ 1/4 \ 1/4 \ 1/4 \ 1/4)$

$(\ 1 \ -1 \ 0 \ 0)$

$(\ 0 \ 1 \ -1 \ 0)$

$(\ 0 \ 0 \ 1 \ -1)$

第 5 章 回归分析的其他应用实例

第 5 章借助于示范性的 SPSS 分析介绍了回归分析方法的其他用途。

第 5.1 节阐述了两种形式的偏回归。第 5.1.1 节介绍了部分最小平方回归（Partial Least Squares, PLS）。尤其是在有很多预测变量、预测变量高度相关，并且/或者预测变量的数量超过个案的数量时，就可以使用 PLS 回归。PLS 回归兼具主成分分析和多元回归的特点，从而可以将任意测量水平的、任意数量的（潜）变量之间的因果关系模拟成线性的结构化方程模型。此外，PLS 还支持混合回归模型和混合分类模型。自变量和因变量既可以是定距的，也可以是定类的。从 SPSS 16 版开始提供了 PLS 命令。PLS 是基于 SPSS 的 Python 扩展。在第 5.1.2 节介绍了利用 SPSS 过程命令 REGRESSION 进行相关分析的一种偏回归形式。

第 5.2 节介绍了如何利用线性混合模型（SPSS 过程命令 MIXED）对个体生长曲线进行线性建模。个体生长建模（individual growth modeling）大致上可以改写为“对个体进行重复测量的方差分析”。对于“普通”线性回归而言，只有一条回归线（例如，回归线也会利用轮廓图生成重复测量的方差分析）通常不适合不同个体（线性）的运行曲线。但是在进行重复测量的回归分析或者方差分析之前，利用随机截距模型进行建模，就可以根据截距、斜率和两个参数同时估计出个体的运行曲线。借助于一个分为三阶段的实例分析，下文演示了某个培训项目所有被试者的成绩在经过一段时间后是否以及在多大程度上有区别。在这个实例中具体检验了：（a）被试者的（成绩）水平是否波动（截距），（b）被试者成绩提高的幅度和速度是否不同（斜率），以及（c）在考虑到被试者（成绩）水平的情况下，他们成绩的提高幅度是否不同（两个参数）？

第 5.3 节介绍了岭回归（SPSS 宏“Ridge-Regression.sps”）。岭回归可以（主要是通过目视）检验的是，可能具有多重共线性的数据是否可以用多元线性回归分析来进行分析。与其他统计方法相反，SPSS 岭回归没有采用菜单导航，而是只能采用宏的形式。但是，岭回归的实

施并不复杂。本章主要演示了多重共线性的可视化，以及如何针对所选择的 K 值计算岭回归。由于 2008 年发行的 SPSS 16 版没有宏 “Ridge-Regression.sps”，因此这里的实例主要是基于 SPSS 15 版的宏。

5.1 偏回归

本章介绍偏回归（又称部分最小平方回归，Partial Least Squares, PLS）（参见 Vinzi et al, 2008、Cohen et al., 2003³、Wentzell & Vega, 2003、Hulland, 1999、Wold, 1985, 1981 和 Pedhazur, 1982²）。PLS 回归利用部分最小平方对回归模型做出估计，也可称为隐结构投影（Projection to Latent Structure）。偏回归是一种真正的多变量分析法。因此，只有当模型中出现两个或多个预测变量和/或因变量时，方可进行偏回归分析。

PLS 部分最小平方回归最初由 Herman Wold（1981, 1985）提出，用于经济学领域，但是不久后就拓展至化学、医药以及市场营销等研究领域。从根本来看，PLS 是一种十分普遍的方法，用于对任意测量水平的、任意数量的（潜）自/因变量之间的因果关系进行建模，即建立结构方程模型。由此可见，在 SPSS 中实现的是线性 PLS 回归（又称偏回归）的一种特殊情况。例如，因变量是定比的。如果因变量是二分的，则该方法也被称为线性 PLS 判别分析。

例如，在下列情况下，PLS 回归可用作 OLS 回归的替代方法。

- 当存在多个预测变量（在此种情况下多重共线性的概率增大，并使得 OLS 回归分析的可靠性下降。如果预测变量的数量由于多重共线性而减少，则可能需要人工分析）时。
- 当预测变量高度相关（多重共线性）时。
- 当预测变量的数量超过个案的数量（OLS 回归分析一般是以过拟合（“完美”）的模型对此作出反应，但是根据经验，这种模型无法通过交叉确认检验）时。
- 当采用探索性研究方法（例如，在使用复杂的回归模型或线性结构方程模型之前），并且注重预测而不是（首先）注重解释时。

PLS 回归兼具主成分分析和多元回归分析的特点。PLS 回归首先进行主成分分析步骤，提取一组可以尽量解释自变量和因变量之间协方差的潜在因子，然后进行回归步骤，即用测定的因子来预测因变量的值。

由于 PLS 回归将主成分分析和回归分析结合起来，因此可以测定在最大程度上解释自变量和因变量之间协方差的一组潜在因子。因此，潜在因子即为所观察到的自变量的线性组合。基于包含 Y 变量在内的叉积，以这样的方式测定独立的潜在因子，一方面明确表明 PLS 回归是一种对预测做出选择的方法，但另一方面也使得对载荷进行解释的难度加大（这是由于模型中的所有变量被额外地居中和标准化，见下文）。预测变量（ X 变量）和因变量（ Y 变量）一样，通过主成分分析被分解为分量。然后重新利用 X 分量预测 Y 值。此时（迭代地）测定 X 主成分，从而将每个 X 值与 Y 变量的协方差实现最大化。

由上述主成分分析可得出：（1）原来的预测变量之间（可能）的多重共线性并不显著，而测定的 X 分量最终垂直于 Y 的预测变量；（2）由于多个变量减少为很少几个主成分，所以变量的原先数量（以及变量/个案数量的比例）不再重要。利用回归分析（当因变量是定比的

时)再次使用 Y 变量的预测值, 就可以最佳地对 Y 变量的观察值做出预测。

PLS 回归这样就可以用于处理多个变量、多重共线性以及变量比个案数量多的情况。因此 PLS 回归优先用于预测, 但是不适用于解释。其原因在于: PLS 回归无法把不影响因果的变量从模型中排除 (Tobias, 1997, 1)。经过细致观察, 借助 PLS 回归只能从表面上解决多重共线性问题。基于主成分分析法, 尽管多重共线性并不影响主成分的测定, 但却影响其解释。由于存在多重共线性, 因此变量影响到多个因子 (即交叉载荷)。多重共线性在预测变量中的影响越深, 就越难产生一个简单的因子结构并对其做出解释。

相较于主成分回归 (principal components regression, PCR; 当时 SPSS 还不支持该方法), PLS 回归法毫不逊色, 甚至可以说是更具优势。PLS 回归在大多数情况下都能够得出更精确的预测变量, 而且比 PCR 分析的参数需求量小得多。因此 PLS 所使用的潜变量也少于 PCR。但是 PCR 分析的潜在因子更易于解释 (参见 Wentzell & Vega, 2003, 257)。

PLS 的假设基本上采用了多元线性回归分析的所有假设 (例如, 线性、无离群值等), 但不包含多重共线性以及显著性检验这两个核心假设。由于线性 PLS 回归是一种没有分布的方法 (因此 PLS 的分布是未知的), 所以无法进行常规的显著性检验。由于 SPSS 目前还没有作为替代的 Bootstrap 法, 因此在分析中, 需注意观察其他回归分析的其他条件。

SPSS 自 16 版起即可用 PLS 过程计算线性偏回归 (参见第 5.1.1 节), 尚不支持非线性偏回归 (NLPLS)。在第 5.1.2 节中介绍了针对 SPSS 较旧版本的另一种利用 REGRESSION 过程的偏回归方法。PLS 可以对单变量和多变量模型做出估计。当有一个或多个定距因变量时, 估计出一个回归模型; 当有一个或多个定类因变量时, 估计出一个分类模型。此外, PLS 还支持混合回归模型和混合分类模型。自变量 (预测变量) 同样有可能是定距或者定类的。字符串变量被自动估计为模型中的分类变量。

5.1.1 运用 PLS 过程 (Python Extension) 进行计算

前提条件和预设置

PLS 和 EXTENSION 命令自 SPSS 16 版之后才得以应用。旧版本 SPSS 的用户在计算偏回归时请参见第 5.1.2 节。必须按以下顺序准确和完整地安装好下面所列的附加程序。

- Python 2.5、Numpy 1.0.1、Scipy 0.5.2。
- SPSS-Python 集成插件 16.0。
- SPSS-PLS 扩展模块 16.0。
- 附加模块: SPSSAUX、SPSSDATA、EXTENSION 和 NAMEDTUPLE (后面的模块可直接存储在创建好的文件夹中, 例如, “C:\Python 2.5”)。PLS 程序 “plscommand.xml” 必须存在, 并且知道其存储位置。

由此可能会产生一个困难, 即针对现有的操作系统, 如何从 <http://www.spss.com/devcentral> 提供的不同程序版本中做出正确的选择。必要时可能需要进行多次安装尝试, 直到 PLS 正常运行为止。在“帮助”和“信息”栏中查看现有 SPSS 版本是否是最新版, 可能会有所帮助。可以查看 SPSS 标志的右边的一个提示信息, 如“16.0.1 (07.12.2007) 版本”。这条信息对正确地选择适合版本的程序和模块来说非常重要。

语句:

```
get file= "C:\...\CP193.sav" .

EXTENSION
    ACTION=ADD
/SPECIFICATION COMMAND=
    "C:\Programme\SPSSInc\SPSS16DE\extensions\plscommand.xml" .

PLS Importe WITH InProd Lager Konsum
/ID VARIABLE = Jahr
/MODEL InProd Lager Konsum
/OUTDATASET CASES=CP193_PLS
    LATENTFACTORS=CP193_LFK
    PREDICTORS=CP193_PRD
/CRITERIA LATENTFACTORS=3.
```

语句说明

通过 GET FILE 语句调用一个 SPSS 数据集 CP193.sav。这个数据集是基于 Chatterjee & Price (1995², 193) 的著作中关于进口额数据的例子。

EXTENSION 命令

从 SPSS 16 版开始提供的 EXTENSION 命令扩展了 SPSS 的功能（即所谓的扩展命令，如 PLS 命令），这些功能是利用外部编程语言（如 Python 编程语言）编写的。扩展命令以及用外部编程语言编写的功能可以由用户自己编写，或者由其他供应商（SPSS, Python）提供。

只有当包含 PLS 专用语句的特定 XML 文件（如 plscommand.xml）通过 EXTENSION 命令和 ACTION=ADD 命令传递给 SPSS 系统时，系统才会识别扩展命令（如 PLS 命令）。通过使用 SPECIFICATION COMMAND 命令将定义扩展命令（如 PLS 命令）专用语句的 XML 文件名称和存储地址（如“plscommand.xml”，存储目录“C:\...”）告知系统，就可通过 EXTENSION 命令将这个扩展命令传递给 SPSS。通过 ACTION=ADD 命令将 XML 文件“plscommand.xml”传递给 SPSS（通过 REMOVE 命令可以删除这个扩展命令）。这样，即可通过 EXTENSION 命令将“普通”SPSS 语句的接口扩展为外部编程语言的功能了。

PLS 命令

PLS 命令给出了计算偏回归的语句。PLS 命令读取（已打开的）活动数据集并执行相应的处理命令。在要执行 PLS 命令的系统上，必须事先安装好 Python 扩展模块“SPSS-PLS Extension Module”（SPSS-PLS 扩展模块）。

备注：PLS 扩展模块取决于 Python 软件，而 SPSS 不是 Python 软件的所有人或者许可证签发人。所有 Python 用户都必须遵守 Python 网站上的 Python 许可证协议规定。SPSS 对 Python 程序的性能不做任何解释，对使用 Python 程序的有关事宜概不负责。

PLS 命令进行偏回归的计算。根据 PLS 命令直接给定（定距）因变量（“进口额”），根据 WITH 命令给定（定距型）协变量，根据 BY 命令给定因子。由于数据集中的变量已经在“测量水平”（在“变量视图”一项下）一项下是定距的，因此通过 MLEVEL 命令（不在本

例中) 定义或改动尺度水平就显得多余了。

通过 MLEVEL 命令能够明确地定义模型中自变量和因变量的尺度水平。当因变量运行 MLEVEL=S (“scale”) 命令时, 自动估计出一个回归模型。当因变量运行 MLEVEL=N (“nominal”) 命令时, 自动估计出一个分类模型。尽管可以利用 MLEVEL=O 命令 (“ordinal”) 对变量进行定义, 但目前只能将其分析为分类变量。只要有一个定类因变量 (不在本例中), PLS 命令利用 REFERENCE 命令和选项 (FIRST、LAST 及数值) 就可以给出参考类别, 以便对参数进行估计。

根据/ID VARIABLE= 命令, 可以为个案输出和存储估计值给定一个标识变量 (字符串, 定量的) (例如 “年”)。

根据/MODEL 命令可以给定模型效应, 主效应通过简单地给定预测变量来定义, 例如 “InProd Lager Konsum” (参见前面的例子)。主效应之间的交互作用可以用 “InProd*Lager” 类别来予以定义 (不在本例中)。PLS 命令不支持嵌套式类别。

根据/OUTDATASET 命令, 可以将模型估计值存储在 (给定的) 数据集中, 并通过图形显示出来。PLS 的预设置使其既不能存储也不能显示模型估计值。利用 CASES 命令将个案的下述估计值存储在给定的数据集 CP193_PLS 中: 预测值、残差、与潜在因子模型的距离、潜在因子值。CASES 命令也输出潜在因子值的图形。

利用 LATENTFACTORS 命令, 将潜在因子权重和潜在因子载荷存储并显示在给定的数据集 CP193_LFK 中, LATENTFACTORS 命令也输出潜在因子权重的图形。利用 PREDICTORS 命令, 将回归参数和 VIP 值 (变量投影重要性指标) 的估计值存储并显示在给定的数据集 CP193_PRD 中, PREDICTORS 命令也输出每个因子的 VIP 值的图形。

根据/CRITERIA 命令, 可以在 LATENTFACTORS= 命令后面给待提取的潜在因子设定一个数量上限。在本例中, 最多只允许提取三个潜在因子。实际提取的因子数量可能与预设置的数量不同。始终以理论为依据给定待提取因子的数量。

PLS 命令会从分析中删除用户定义的和系统定义的缺失值。现在 PLS 既不能进行交叉验证, 也不能根据拟合优度评估模型拟合优度。

输出结果

PLS 回归

PLS 过程的输出结果现在还只能使用英语。

解释方差的比例

潜在因子	统计量				
	X 方差	累积 X 方差	Y 方差	累积 Y 方差	(R ²) 调整 R ²
1	0.693	0.693	0.968		0.968 0.966
2	0.306	1.000	0.005		0.973 0.969
3	0.000	1.000	0.000		0.973 0.967

表“解释方差的比例”反映了通过潜在因子解释的方差在自变量和因变量中所占的比例。理想情况下，一个模型可以对 X 变量和 Y 变量进行最大程度地解释。而就实际情况而言，有时可以对 X 变量进行更多解释，有时可以对 Y 变量进行更多解释。在 CRITERIA 一项下，根据 LATENTFACTORS 命令最多可以调用三个潜在因子，每个潜在因子（和每行）都会给出一个解释方差。

第一个潜在因子（“1”）解释了约 70% 的预测变量方差（“ X 方差”）和约 97% 的因变量方差（“ Y 方差”）。第二个潜在因子（“2”）解释了约 30% 的预测变量方差（“ X 方差”）和约 0.5% 的因变量方差（“ Y 方差”）。前两个潜在因子总共解释了 100% 的预测变量方差（“累积 X 方差”）和约 97% 的因变量方差（“累积 Y 方差（ R^2 ）”和调整 R^2 ）。第三个潜在因子解释了约 0% 的预测变量方差（“ X 方差”）和约 0% 的因变量方差（“ Y 方差”）。所有三个潜在因子总共解释了 100% 的预测变量方差（“累积 X 方差”）和约 97% 的因变量方差（“累积 Y 方差（ R^2 ）”）。第三个潜在因子不再起到解释预测变量方差和因变量方差的作用，因此可以从模型中删除。在本例中，利用两个潜在因子的解决方案就已经能接近最佳地解释 X 变量和 Y 变量（分别 100% 和 97%）。既然已经能解释 97% 的因变量方差和 70% 的预测变量方差，那么只要在提出的问题中因变量的方差解释优于自变量的方差解释，有时候出于减少参数需求量的原因甚至可以考虑用仅有一个潜在因子的模型。

一个潜在因子对 Y 变量方差解释得越多，那么它就越能够解释非独立数据另一个样本的数据方差。而一个潜在因子对 X 变量方差解释得越多，那么它越能够描述自变量当前样本的观察值。

参数	
自变量	因变量
	进口额
（常数）	-19.725
国内生产总值	.032
库存	.414
消费	.243

表“参数”包含了用于预测因变量“Importe”（“进口额”）的预测变量（“InProd”国内生产总值、“Lager”库存、“Konsum”消费）回归系数的标准化估计值。对系数的解释相当于对普通回归系数的解释。因此，数值（效应的重要性）和正负号（效应的方向）是很重要的。所有预测变量都与“进口额”正相关。“库存”（0.414）和“消费”（0.243）的系数明显高于“国内生产总值”（0.032）的系数。就这方面来说，对于预测进口额数值的重要性更大。

如果是定类因变量，则 SPSS 输出结果基本上相当于逻辑回归的输出结果（不在本例中）。

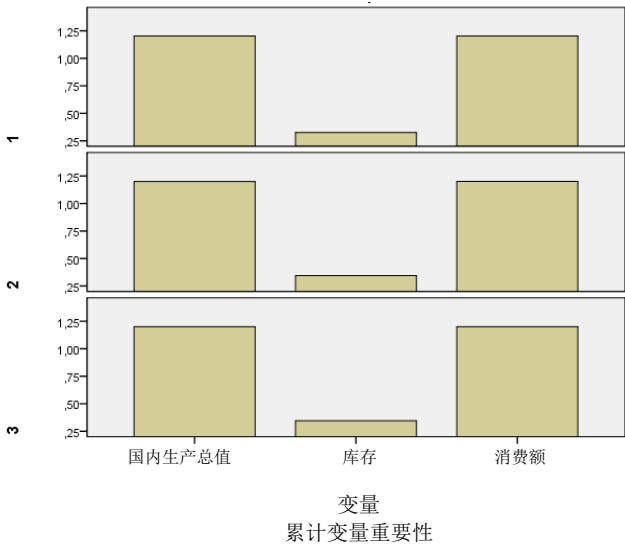
变量投影重要性			
变量	潜在因子		
	1	2	3
国内生产总值	1.203	1.200	1.200

续表

变量	潜在因子		
	1	2	3
库存	.325	.345	.345
消费	1.203	1.201	1.201

累积变量重要性。

表“变量投影重要性”（“VIP”）反映了在预测模型（“投影”）中每个自变量对于每个潜在因子的重要性（参见下图“累积变量重要性”）。VIP 系数表明了同时模拟 X 得分和 Y 得分（见下文）时每个预测变量的重要性。因此“国内生产总值”和“消费”（分别都是 1.20）明显具有比“库存”（0.32）更大的重要性。在超出所测定潜在因子的范围时，预测变量的重要性基本上保持恒定。根据 Wold（1994）的观点，每个 VIP 系数小于 0.8、同时回归系数也很小的变量（参见上文“参数”表）都能从模型中剔除。预测变量“国内生产总值”、“消费”和“库存”并不满足这两个条件，因此仍保留在模型中。



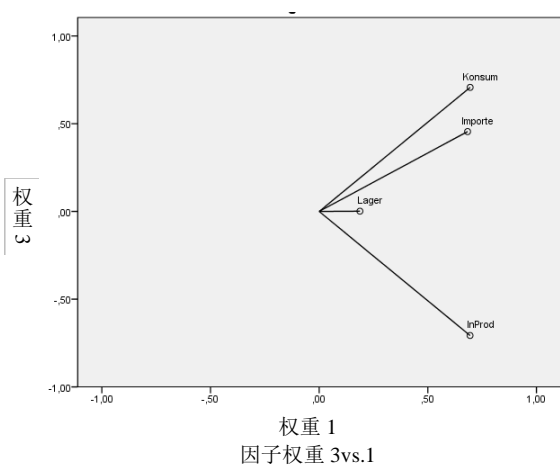
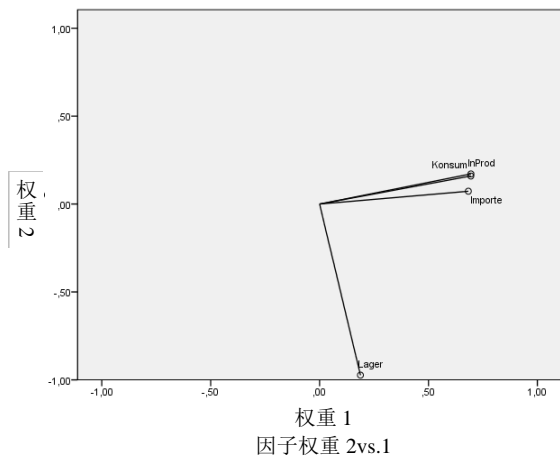
图“累积变量重要性”反映了上文表“变量投影重要性”（“VIP”）中的数值。由此可以看出，“国内生产总值”和“消费额”具有比“库存”更大的重要性，并且几乎恒定不变。“国内生产总值”和“消费额”的重要性几乎相当，而与之相比，“库存”的数值要低很多，其重要性也就相去甚远了。

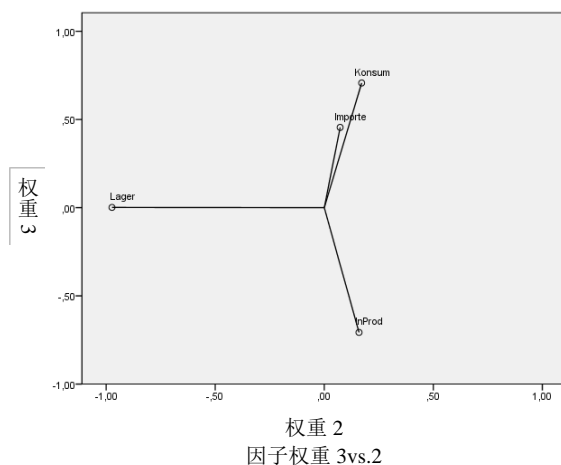
权重				载荷			
变量	潜在因子			变量	潜在因子		
	1	2	3		1	2	3
国内生产总值	0.694	0.159	-0.707	国内生产总值	0.688	0.132	-0.708
库存	0.188	-0.973	0.001	库存	0.234	-0.982	0.008
消费	0.695	0.172	0.707	消费	0.688	0.134	0.706
进口额	0.683	0.073	0.455	进口额	1.000	1.000	1.000

表“权重”和“载荷”反映了因变量或者自变量在所测定潜在因子中的权重和载荷。权重和载荷表明，每个自变量对各自的潜在因子起到多大作用。权重代表 X 变量与 Y 值的相关，载荷则代表每个 X 变量的重要性，常用作对因子的命名。然而，载荷并不总是容易解释的。例如，当存在交叉载荷时，即一个变量对多个因子产生显著载荷时。交叉载荷大多是由多重共线性引起的。预测变量中的多重共线性越明显，就越难建立简单的因子结构，对其解释也就越加困难。根据 Hulland（1999，198）的观点，验证性 PLS 中的载荷在理想情况下应该至少达到 0.7，以确保一个因子能代表一个自变量。Raubenheimer（2004）将探索性 PLS 的分界值设定为 0.4。可以试验性地从模型中删除权重和载荷低的预测变量，因为这样有可能更好地解释 Y 方差。

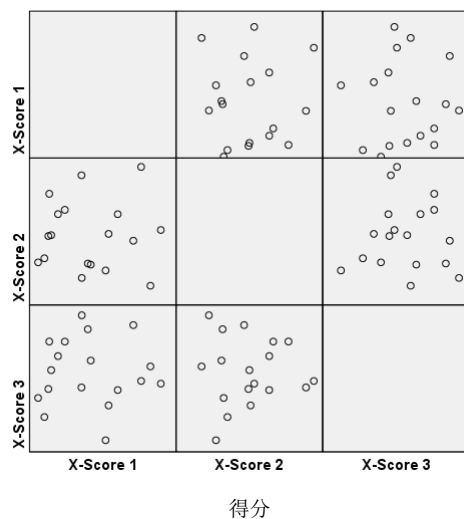
表“权重”表明自变量“国内生产总值”与第一个因子呈高度的正相关，并且与第三个因子呈负相关。“库存”仅与第二个因子呈高度的负相关。“消费”与第一个因子和第三个因子均呈高度的正相关。下面的“因子权重”图形反映了表格“权重”中的因子权重。

表“载荷”表明自变量“国内生产总值”在第一个因子上具有高度的正载荷，在第三个因子上具有高度的负载荷。“库存”仅在第二个因子上具有高度的负载荷。“消费”在第一个和第三个潜在因子上的载荷显著，所以“消费”就是一个“交叉载荷”变量。与因子分析相似，载荷的解释始终是以理论为基础的，从来不是仅根据随机选择的极限值。在本例中，没有将任何预测变量从模型中删除，因为它们的数值都高于分界值， Y 方差的解释是最佳的。





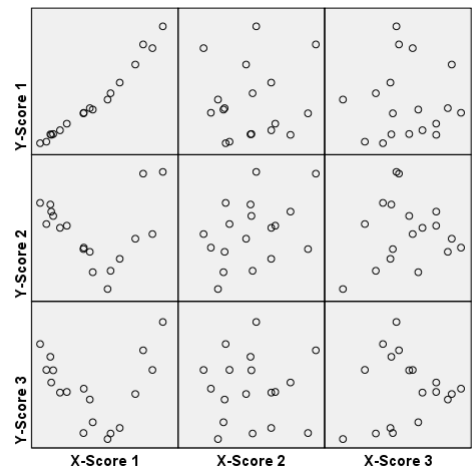
系列图“因子权重”可视化了表“权重”中的权重。因为一个权重相当于一个相关性，相关性的数值（也就是与中心的距离）则表明了每个自变量对各自潜在因子的贡献率。由此看来，靠近交点的变量对模型的贡献很小，可以试验性地从模型中删除。一个变量离交点越远，它对模型的贡献就越大。在这组图形中，所有的变量都远离交点，因此没有从模型中删除预测变量。



图“得分”反映了各个潜在因子的 X 得分相对于其他潜在因子的 X 得分的情况。例如，单元格“X-Score 1”和“X-Score 2”截取了潜在因子 1 和 2 的 X 值。图“得分”没有展示所观察预测变量的数值。从“得分”图中可以看出，这里不存在特殊的分布格局。

图“回归图： X 得分 vs. Y 得分”以回归图的形式展示了三个潜在因子的 X 值和 Y 值（因为调用了三个潜在因子，所以图形中显示了三个）。在 X 轴截取各潜在因子的 X 值，在 Y 轴截取各潜在因子的 Y 值。从这个图形中可以看出，在每个潜在因子内部， X 值与 Y 值之间关联的类型、程度和方向。例如，“X-Score 1”和“Y-Score 1”一行截取了潜在因子 1 的 X 值和 Y 值。从这个图中还可以看出，在这里存在特殊的分布格局。第一个潜在因子 X 值与 Y 值之间的关联明显是线性的。这个调查结果表明，模型是稳定的。与之相反，第二个潜在因子 X 值与 Y 值之

间的关联明显是云状的。这个发现表明模型是弱的。



回归图：X 得分 vs.Y 得分

图“得分”并没有展现三个不同的模型，而是展现了一个模型内部所调用潜在因子的估计 X 值和 Y 值。

此外，用户还可以利用可视化来检验存储在数据集 CP193_PLS 中的残差的离群值，非正交性和异方差性（不再继续进行残差分析）。

5.1.2 运用 SPSS 过程 REGRESSION 进行计算

这种分析方法的名称“偏回归”（又称偏回归分析）源自于：一次回归含有两个或多个预测变量，在这些预测变量中分别根据其中两个变量测定回归系数或相关系数，而这两个变量事先分别已经剔除了其他预测变量的方差分量（参见 Cohen 等人著作，2003³，69-75、Litz，2002，第 3 章）。因此，可以在预测变量彼此相关，即存在多重共线性的情况下使用偏回归。通过这种形式的偏回归，可以查明一些变量（例如，a、b 和 c）中的单个变量（如 c）是否能影响所调查的 x（预测变量）与 y（标准）之间关联。

与 PLS 相反（参见第 5.1.1 节），这种形式的偏回归是以相关分析或者回归分析为基础的，因此用 SPSS 过程命令 REGRESSION 来进行计算。

下面的语句实例调查了，一个或若干个变量（例如，a、b 或 c）是否能影响我们所感兴趣的 x 与 y 之间关联。对此所需的处理方法以 Litz（2002，77-91）的著作为准。这里不再根据 REGRESSION 输出结果或者 PARTIAL CORR 输出结果进行示范性的解释，因为这种调查的前提条件是已经知道了对其的解释。

下列后续步骤是必要的。

（1）对双变量初始模型进行回归分析。

计算 x 与 y 之间的回归，无须考虑更多变量可能的（干扰）效应。这是为了测定模型的参考参数所必需的步骤。

（2）计算零阶双变量的相关性。

通过用变量 a 、 b 和 c 计算 x 与 y 的相关性， a 、 b 和 c 可能会对 x 与 y 的相关性产生（干扰，干涉）预测。这个步骤对于估计多重共线性的程度是必要的。下一步是要对与 x 或 y 显著相关的变量 a 、 b 或 c 进行更进一步的观察。

（3）测定模型的偏相关。

计算模型的偏相关，一种方法是一次性计算变量 a 、 b 和 c 与 x 和 y 的偏相关，另一种方法是逐个地分别计算。这个步骤对于估算偏离程度（去除“外部方差”）而言是必要的。将显著预测 x 与 y 之间关联的变量，如（假定为） c ，纳入接下来的计算步骤。通过调整相关系数和偏相关系数表中 x 与 y 之间的系数，即可看出这个点。如果 x 与 y 之间关联的相关系数发生了明显变化，则在计算时要将与此有关的变量从关联中“排除”。

（4）利用所选择的干扰变量作为预测变量进行回归分析。

以 x 和 y 分别作为标准（和初始模型中的预测变量），以干扰变量 c 作为预测变量，分开两组进行回归分析计算。这个步骤对于从 x 和 y 中剔除 c 的预测是必要的。分别保存非标准化残差。

（5）进行偏回归分析。

以步骤（4）中的非标准化残差，而不是以变量 x 和 y 为基础进行回归分析。

这个系列分析步骤仍需补充（图形式）残差分析。对此，线性回归分析和线性相关分析的所有前提条件都适用。此外，不允许出现没有将干扰变量从模型中删除的情况，因为这些干扰变量人为地表明存在一种原先假定的关联，但实际上这种关联并未出现。

（1）对双变量初始模型的回归分析。

```
REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN (.05) POUT (.10)
/NOORIGIN
/DEPENDENT y
/METHOD=ENTER x
/SAVE RESID
```

（2）计算零阶双变量的相关性。

```
CORRELATIONS
/VARIABLES= x y BY a b c
/PRINT=TWOTAIL NOSIG
/MISSING=LISTWISE
```

（3）测定模型的偏相关。

```
PARTIAL CORR
/VARIABLES= x y BY a b c
/SIGNIFICANCE=TWOTAIL
```

```
/STATISTICS=DESCRIPTIVES CORR  
/MISSING=LISTWISE
```

PARTIAL CORR

```
/VARIABLES= x y BY a  
/SIGNIFICANCE=TWOTAIL  
/MISSING=LISTWISE
```

PARTIAL CORR

```
/VARIABLES= x y BY b  
/SIGNIFICANCE=TWOTAIL  
/MISSING=LISTWISE
```

PARTIAL CORR

```
/VARIABLES= x y BY c  
/SIGNIFICANCE=TWOTAIL  
/MISSING=LISTWISE
```

(4) 利用干涉变量作为预测变量进行回归分析。

REGRESSION

```
/MISSING LISTWISE  
/STATISTICS COEFF OUTS RANOVA  
/CRITERIA=PIN (.05) POUT (.10)  
/NOORIGIN  
/DEPENDENT x  
/METHOD=ENTER c  
/SAVE RESID
```

REGRESSION

```
/MISSING LISTWISE  
/STATISTICS COEFF OUTS RANOVA  
/CRITERIA=PIN (.05) POUT (.10)  
/NOORIGIN  
/DEPENDENT y  
/METHOD=ENTER c  
/SAVE RESID
```

(5) 进行偏回归分析。

REGRESSION

```
/MISSING LISTWISE  
/STATISTICS COEFF OUTS RANOVA  
/CRITERIA=PIN (.05) POUT (.10)  
/NOORIGIN  
/DEPENDENT res_1  
/METHOD=ENTER res_2
```

5.2 个体生长曲线

回归分析也可以用于对重复测量的分析（例如，短时间序列）。这个用途被称为个体生长曲线建模（individual growth modeling）。由于这种方法将个体的回归系数用作随机变量，因而也被称为随机截距模型。这种方法假设，回归系数是正态分布的。在下面这个例子中，仅计算线性模型。

通过用时间序列分析方法和重复测量的方差分析方法来调整个体生长曲线建模，并根据所表现出的差异和共同点大致上阐明个体生长曲线建模的优点和缺点。

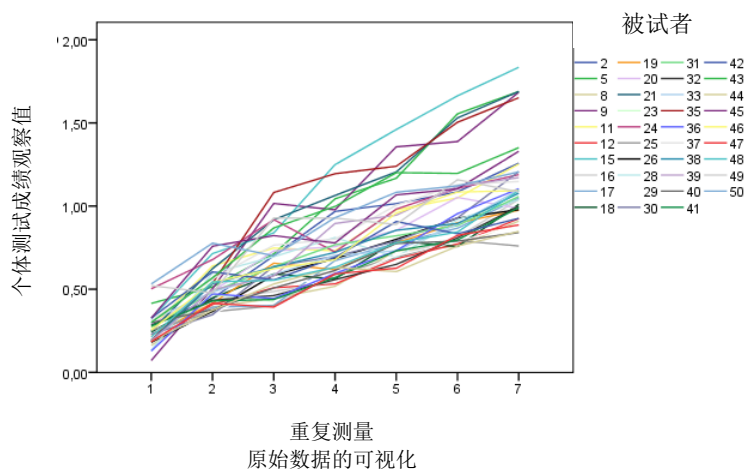
与时间序列分析方法和重复测量方差分析方法的共同点在于，所有这三个方法都是基于相关的测量值序列。在这些相关的测量值序列中，每个测量时间点均与一个测量值相对应。与时间序列分析方法的不同之处在于：在采用个体生长曲线时，允许所调查的时间序列要短得多（例如，测量时间点数量自 $N=3$ 起）。此外，个体生长曲线建模是一种回归分析方法，并不能对真正的时间序列分析效应，如趋势、周期、季节性等进行建模。

与重复测量的方差分析的差异在于：重复测量的方差分析将原本个体的重复测量整合为群组。在对个体生长曲线建模时，图中的一条线就反映了一个个体的数值。但是，对重复测量的方差分析进行可视化（如轮廓图），就需用直线反映一组个案的概括数值。下节将对这个差异进行探讨。首先探讨的是最后一个特点，即前面所介绍三种个体生长曲线建模方法的特点。对于曲线性数值序列，这些方法则不太适用。

回到对个体生长曲线（不论线性或曲线性）的总合：这种概括并不总是适合的，例如，在（多个）个体的走向趋势既有向上又有向下时。适用于所有个案的线性（或者非线性）回归直线给人的感觉是，就好像这些截然不同的趋势根本不存在一样。而总和则使人们看不到这些截然不同的趋势（例如，在进行重复测量的方差分析时所使用的轮廓图中）。但是在对个体生长曲线建模时，将每个个案的走向可视化为直线，并进行分析。如需研究（多个）个体的走向趋势是否一致，或者这些个体是否可能具有不同的线性（或者非线性）走向模式，则这种分析方法就是合适的。

因此，在进行重复测量的方差分析之前，建议对个体生长曲线进行建模。因为只有如此才能检验出，用于重复测量的方差分析的个案或个案组是否具有同质的走向趋势。如果有异质的走向趋势，则有两种方法将可能存在的误差方差控制到最小：将具有异质走向趋势的个案组进行分组，从而使所产生的个案子组具有同质的走向趋势。未分组的数据可以划分到具有某些走向模式的个案组中。

下图示例中所用数据来自作者的一项调查。在这项调查中，39 个人连续 7 天参加同样的神经心理学培训。该培训的目的是，提高被试者的神经心理表现。最后得出 39 人连续 7 次重复测试的成绩，每个人都有 7 个测试值。从图形可以看出，个体走向是接近于线性的数值序列。下图显示了 39 个被试者的成绩变化曲线。后续的分析步骤旨在分三步检验培训是否成功，以及各个患者以何种方式改善其神经心理表现。



如上图所示，神经心理培训效果显著，大多数被试者成绩持续提高。但是也可以看到，个别的成绩曲线是完全异质的。例如，相较于被试者 47，被试者 5 的神经心理表现提高幅度呈现显著加快的趋势。只有一条回归线（例如，只有一条回归线将创建包括重复测量的方差分析）就无法适应不同的趋势。因此，基于随机截距模型对个体生长曲线建模就是显而易见的应采取的方法。

后续分析分为以下三步。

- 第 5.2.1 节检验了是否每个个案都具有不同的截距。换言之，是否每个被试者都具有不同的成绩水平？
- 第 5.2.2 节检验了是否每个个案都具有不同斜率。换言之，是否被试者的成绩以不同的程度和速度提高？
- 第 5.2.3 节检验了是否每个个案都具有不同的截距和斜率。换言之，在考虑到被试者成绩水平（截距，方法 1）的情况下，他们的成绩是否有不同程度的提高（斜率，方法 2）？

此外，同前两种方法一样，要求最高的第三种方法在原则上可以根据相似的生长曲线将个案分组（最后一种方法对此做了示例性尝试）。但是在对个体生长曲线进行可视化时，往往不容易将各个直线相互区分开来，从而评估个体的表现。由于是对个体生长曲线进行可视化和分析，所以这种方法有时并不适用于大量数据的个案（例如，在 SPSS 演示数据集“testmarket.sav”中）。

对这些方法的预测值进行可视化处理后，通常要通过对原始数据的可视化予以调整。这样做的最主要原因在于，估计过程以及由此得出的估计值可能不是最佳的，有时甚至是无效的。例如，可能发生这样的情况：尽管满足所有的收敛标准，但最终的海森（Hessian）矩阵不是正定的。而随机截距法和随机斜率法的效果很小，因此，从方法上来看，在有些情况下，无法对数据变异做出完整解释。

5.2.1 方法 1：随机截距模型

第一种方法调查了个体成绩的变化过程（随机截距法），它假设每个个案都具有不同截距。

换言之，每名被试者的（成绩）水平都有所波动或提高。因此，调查只针对其（成绩）水平，并不针对其斜率。该模型假设，这些截距呈现平均值为零、方差未知的独立同分布的正态分布。

方法 1 语句：

```
MIXED Y WITH zeit
/FIXED intercept zeit
/METHOD=REML
/RANDOM intercept | SUBJECT (Probandn) COVTYPE (ID)
/PRINT SOLUTION TESTCOV
/SAVE pred (pred_1) .
```

MIXED 语句形式以粗体显示。

方法 1 语句解释。

例如，线型混合模型（Linear Mixed Model, LMM）这样拓展了一般线性模型（GLM）：数据可以是相互关联的，并且具有不恒定的变异性（异方差性）。在这个点上，线型混合模型（LMM）更具灵活性。此种模型设定与 GLM 方法类似：将分类预测变量（也包括重复测量因子，wsfactor）设定为模型的因子，因子的每个变量类别可能都对因变量的相应数值产生不同的线性效应。在这种情况下，必须对确定性效应与随机效应进行区分。

- 变量是确定性因子，其所有分类是已知并且存在的。
- 变量是随机性因子，其现有分类只是随机选择了所有可能的因子变量类别的一部分。

将定比预测变量设定为模型中的协变量，协变量分别与因变量呈线性相关，也就是说，所有可能的因子等级的每个组合方式。此外，LMM 能够设定随机效应的协方差结构。在基本设置时，随机效应是相互不相关的，并且具有相同的方差。

MIXED 命令调用一个线型混合模型，其中 *Y* 为定比因变量、*TIME* 为重复测量因子。按照 **FIXED** 命令，**ZEIT** 与 **INTERCEPT** 命令定义了确定性效应。根据 **METHOD** 命令，给出了估计方法。除了已介绍的受限最大似然法（REML, Restricted maximum likelihood）外，SPSS 还提供最大似然法（ML, Maximum likelihood）。根据 **RANDOM** 命令确定随机效应，由于有 **INTERCEPT** 命令，所以该随机效应由截距项组成，同时各个个案（**PROBAND**）由 **SUBJECT** 命令识别。借助 **ID** 命令，**COVTYPE** 命令定义了类型标识的协方差结构。

细心的读者在这里会注意到，**INTERCEPT** 命令既可在 **FIXED** 命令下给定，也可在 **RANDOM** 命令下给定。有趣的是，即使在删除 **FIXED** 命令下的 **INTERCEPT** 命令时，SPSS 也能测定出一个常数项，并且与 **INTERCEPT** 命令的设定值相比，SPSS 也能得出绝对相同的结果。因此，在 **FIXED** 命令下是否给定 **INTERCEPT** 命令并无区别。

这里接受 SPSS 预设置作为求解算法的选项，当然也可以在 **CRITERIA**（未给定）下单独设置。利用 **PRINT** 命令调用对确定性及随机效应（**SOLUTION**）参数的求解，以及对协方差参数（**TESTCOV**）的检验（尤其是 Wald 检验）。在变量 **PRED_1** 中存储的预测（**PRED**）成绩可通过 **GGRAPH** 线图实现可视化。关于 MIXED 过程的统计及技术细节，请参见 SPSS 16.0 命令语句参考（Command Syntax Reference）（2007）。

结果解释

混合模型分析

模型维度 ^a					
		类别数量	协方差结构	参数数量	被试变量
确定性效应	常数项	1	标识	1	被试者
	时间	1		1	
随机效应	常数项	1		1	
残差				1	
总计		3		4	

a. 因变量：Y。

表“模型维度”可以让我们对设定模型的参数有个大致了解。在模型中，针对每个确定性效应及随机效应都输出了变量类别和参数的数量。对于随机效应，还输出了预设协方差结构（“标识”）的类型及用于识别个案的变量（参见“随机效应”：PROBANDN）。在图例中则给定了因变量“Y”。

信息准则 ^a	
受限制的 2*对数似然值	-266.978
Akaike 信息准则（AIC）	-262.978
Hurvich 与 Tsai（IC）	-262.933
Bozdogan 准则（CAIC）	-253.804
Bayes 准则（BIC）	-255.804

以尽可能小的形式显示信息准则。

a. 因变量：Y。

表“信息准则”给出了对不同混合模型进行比较的标准。基本规则是：数值越小，则模型越好。受限制的 2*对数似然值是选择模型的最简单标准，AIC 基于对数似然值，并利用多个参数对模型进行“修正”。对于少量抽样，AICC 会对 AIC 进行修正，而随着抽样数量越来越多，AICC 将接近于 AIC。BIC 同样也会对过度参数化的模型进行“惩罚”，且严格程度更甚于 AIC。随着抽样数量越来越多，CAIC 将接近于 BIC。在图例中给出了因变量。应事先说明的是，在所有三种随机截距方法中，截距模型的 AIC 值（-262.978）最差（最高）。

确定性效应

确定性效应检验，类型 III ^a				
来源	分子自由度	分母自由度	F 值	显著性
常数项	1	65.032	39.134	0.000
时间	1	229.885	1434.316	0.000

a.自变量：Y。

表“确定性效应检验，类型 III”反映了模型中每个确定性效应的 F 检验。如果某个确定性效应达到较小的显著性值（例如 < 0.05 ），它便会对模型产生预测。在图例中给出了因变量。常数项以及“时间”因子的 p 值都是 $p=0.000$ ，因此它们显著影响模型，然而这并非分析的主要结果。

固定效应的参数估计^a

参数	估计	标准误差	自由度	T 统计量	显著性	置信区间 95%	
						下限	上限
常数项	0.184875	0.029553	65.032	6.256	0.000	0.125854	0.243896
时间	0.138946	0.003669	229.885	37.872	0.000	0.131717	0.146175

a. 因变量：Y。

表“固定效应的参数估计”说明了模型中每个确定性因子（如时间）对于因变量的影响。常数项“估计”一栏中的数值尤为重要。

这个表为模型的每个参数给出了非标准化估计值、标准误差、自由度、T 统计量、显著性以及置信区间。通常只对具有显著性的预测变量的参数，以及置信区间不包括数值 1 的预测变量的参数进行解释。概率则可解释为分别根据其他预测变量做了调整。

“估计”并未标准化，可能有很大的误导作用。

常数项“估计”栏中的数值给出了估计的截距，即整体上的回归线估计（平均）水平。常数项在统计学上是显著的（ $p=0.000$ ），从统计学角度可以认为测试成绩及神经心理表现通常具有不同的水平。随机截距的方差估计值可以从“协方差参数估计”中获取。

协方差参数

协方差参数估计^a

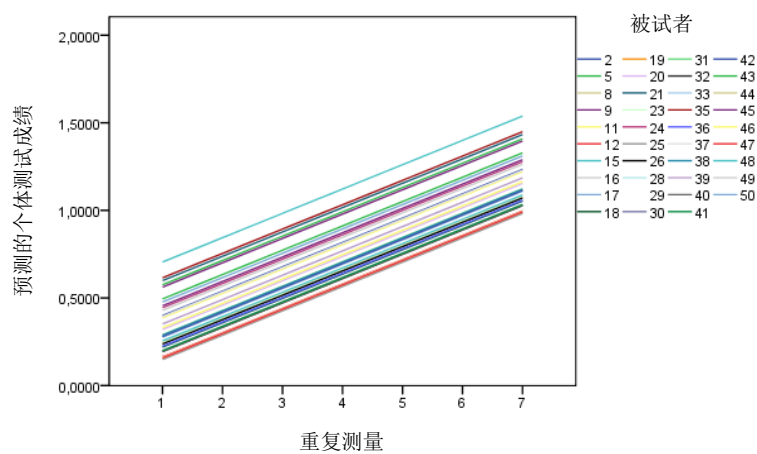
参数	估计	标准误差	Wald 统计量	显著性	置信区间 95%	
					下限	上限
残差	0.014375	0.001343	10.703	0.000	0.011970	0.017264
常数项 [被试者 = probandn 方差]	0.023740	0.005926	4.006	0.000	0.014555	0.038721

a. 因变量：Y。

表“协方差参数估计”总结了用于设定随机效应及残差协方差矩阵的参数，并且针对所计算出的模型输出了非标准化估计值、标准误差、Wald 统计量、显著性以及置信区间。残差与大约 0.14 的方差无关。常数项相当于随机效应 PROBANDN。随机效应具有“标识”类型的协方差矩阵，其方差约为 0.024（参见各自的“估计”）。

“估计”一栏的数据并未标准化，可能有误导作用，因此在对输出值进行解释时需要格外谨慎。根据原始数据的极差以及所获得显著性（ $p=0.000$ ）极差可以得出以下结论：不同的被

试者具有不同的截距。通过对预测的测试成绩进行可视化，证实了这个调查结果。



方法 1：随机性截距的模型

图形解释

这个图形给出了三阶段分析的第一种结果。被试者具有不同的截距（斜率几乎保持恒定）。这意味着它验证了这项调查的设计方案：被试者有不同水平的提高，也就是说，有些被试者在所有七项测试中取得的成绩都比其他被试者高。但是，这个结果也同样表明，被试者的成绩起点可能有所不同。有些被试者的成绩起点高，而有些被试者的成绩起点低。有理由认为他们具有不一致的基线。这个调查结果并没有说明培训起作用的速度，这个问题将在下一个分析步骤进行探讨。

5.2.2 方法 2：随机斜率模型

调查个体成绩变化过程的第二种方法（随机斜率法）是假设每个个案都具有不同的斜率。换言之，不同被试者的成绩以不同速度提高（方法 2 假设其起点相同）。因此，调查对象仅限于被试者成绩的斜率，而非其水平。这个模型假设：这些斜率呈现一种平均值为零、方差未知的独立同分布的正态分布。下面几节只对模型、语句、统计量的新内容或者相关内容进行解释。

方法 2 的语句：

```
MIXED Y WITH zeit
/FIXED intercept zeit
/METHOD=REML
/RANDOM zeit | SUBJECT (probandn) COVTYPE (ID)
/PRINT SOLUTION TESTCOV
/SAVE pred (pred_2) .
```

方法 2 语句的解释

除了两处例外，方法 2 的语句与方法 1 的语句是一致的。这两处区别在于，在 RANDOM 一项下给出了 ZEIT，代替了原先的 INTERCEPT。在 SAVE 一项下，代替原先的变量 PRED_1，将变量 PRED_2 定义为根据方法 2 的成绩预测值的存储变量。通过一个直线图，将存储在 PRED_2 下的成绩预测值可视化。

结果解释

混合模型分析

信息准则^a

受限的 2*对数似然值	-386.559
Akaike 信息准则 (AIC)	-382.559
Hurvich 及 Tsai (IC)	-382.514
Bozdogan 信息准则 (CAIC)	-373.385
Bayes 信息准则 (BIC)	-375.385

以尽可能小的形式显示信息准则。

a. 因变量: Y。

应事先说明的是: 在所有三种随机截距方法中, 斜率模型的 AIC 值是居于第二位的 (-382.559)。只需一个参数, 这个方法就能取得接近方法 3 的数值 (-388.283), 但是还有另外一个参数。可以说, 对模型及其可比性具有决定性的是被试者成绩的斜率, 而不是成绩水平。而且从减少参数需求量的角度来考虑, 纯粹的斜率模型 (方法 2) 比截距斜率模型 (方法 3) 更为优越。与仅仅模拟个体成绩斜率所起作用的方法 2 相比, 纳入成绩水平并没有实质性地改进总体模型。

确定性效应

固定效应的参数估计^a

参数	估计	标准误差	自由度	T 统计量	显著性	95%置信区间	
						下限	上限
常数项	.186215	.012411	230.345	15.004	.000	.161762	.210669
时间	.138245	.006870	50.464	20.124	.000	.124451	.152040

a. 因变量: Y。

在表“固定效应的参数估计”中, “估计”一栏中的数值针对常数项给出了回归线的总体估计 (平均) 斜率。常数项在统计学上的显著效应 ($p=0.000$) 表明, 可以认为测试成绩及神经心理表现通常具有不同的提高幅度。随机斜率的方差估计值可以从表“协方差参数估计”中获取。

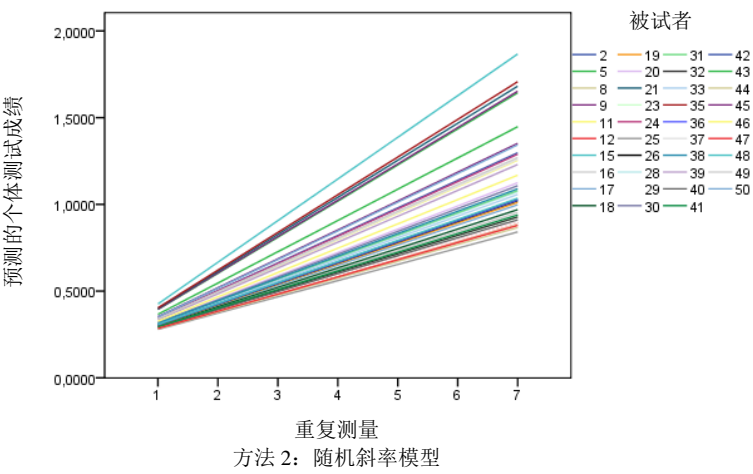
协方差参数

协方差参数估计^a

参数	估计	标准误差	Wald Z	显著性	95%置信区间	
					下限	上限
残差	.008291	.000775	10.702	.000	.006904	.009958
	.001520	.000364	4.172	.000	.000950	.002431

a. 因变量: Y。

表“协方差参数估计”反映了随机效应的估计值。随机效应的方差约为 0.001（参见“估计”一栏）。可以认为，被试者在培训过程中以不同的速度提高其成绩。通过对预测的测试成绩进行可视化，证实了这个调查结果。



这个图形展示了三阶段分析的第二种结果：被试者具有不同的斜率（截距几乎恒定）。这意味着向回推演到这项调查的设计方案：被试者以不同的速度提高成绩，也就是说，有些被试者在所有七项测试中的成绩提高幅度，大于其他被试者。但是这个结果同样表明：这个调查结果并未说明初始值（水平）可能产生的预测。只知道斜率，而不知道初始值的水平（截距）。但是，方法 1 的结果使人们有理由认为被试者具有不一致的基线。在第三个也就是最后一个分析阶段，将会对培训成绩的水平 and 斜率同时进行调查。

5.2.3 方法 3：随机截距和随机斜率模型

调查个体成绩变化过程的第三种方法（随机截距和随机斜率法）假设每个个案都具有不同截距和不同斜率。换言之，不同被试者的成绩在不同的水平上波动，并且以不同的速度提高。与前两种方法不同的是，方法 3 对成绩水平及斜率同时进行调查。这个模型假设，成对的截距与斜率呈现一种平均值为零、协方差未知的 IID 二元正态分布。下面几节只对模型、语句和统计量的新内容或者相关内容进行解释。

方法 3 的语句：

```
MIXED Y WITH zeit
/FIXED intercept zeit
/METHOD=REML
/RANDOM intercept zeit | SUBJECT (probandn) COVTYPE (UN)
/PRINT SOLUTION TESTCOV
/SAVE pred (pred_3) .
```

方法 3 语句的解释

除了三个例外，方法 3 的语句与方法 2 的语句完全一致。这三个区别在于：在 RANDOM 命令下同时给定 INTERCEPT 与 ZEIT。在 COVTYPE 命令下，由 TYP UN（非结构化，完全

通用) 给定协方差结构。在 SAVE 命令下, 将 PRED_3 定义为成绩预测值的存储变量。

结果解释

混合模型分析

信息准则^a

受限 2*对数似然值	-396.283
Akaike 信息准则 (AIC)	-388.283
Hurvich 和 Tsai () (AICC)	-388.131
Bozdogan 准则 (CAIC)	-369.934
Bayes 准则 (BIC)	-373.934

这里显示的信息准则的形式比越小越好。

a. 因变量: Y。

虽然在所有三种随机截距方法中, 截距-斜率模型的 AIC 值最好 (最低, -388.283), 但也只是因为在模型中考虑了两个参数。与纯粹的斜率法相比, 纳入成绩水平并没有实质性地改进总体模型。从内容以及参数经济角度考虑, 纯粹的斜率模型 (方法 2) 比截距-斜率模型 (方法 3) 更为优越。

固定效应的参数估计^a

参数	估计	标准误差	自由度	T 统计量	显著性	95%置信区间	
						下限	上限
常数项	.185831	.016386	38.073	11.341	.000	.152663	.219000
时间	.138434	.006949	37.970	19.921	.000	.124366	.152503

a. 因变量: Y。

在表“固定效应的参数估计”中, 常数项“估计”一栏中的数值给定了在考虑到截距情况下的回归线估计 (平均) 斜率。统计学上的显著性效应 ($p=0.000$) 意味着, 从统计学角度可以认为测试成绩或神经心理表现在不同水平上有不同程度的提高。

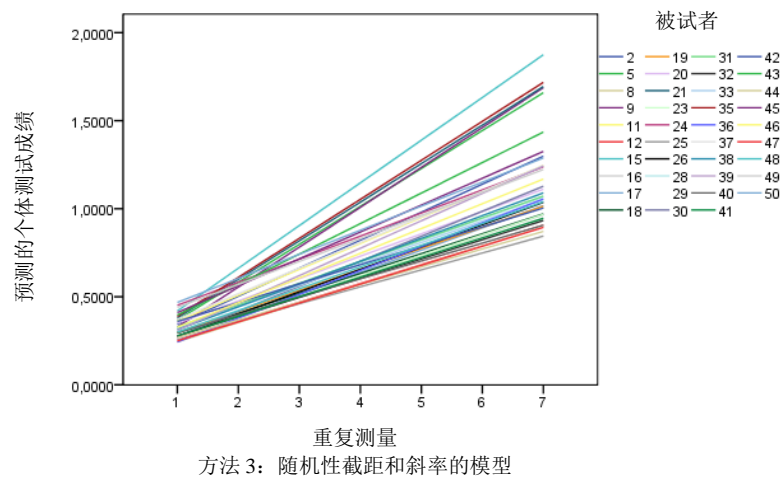
协方差参数

协方差参数估计^a

参数		估计	标准误差	Wald Z	显著性	95%置信区间	
						下限	上限
残差		.007075	.000722	9.795	.000	.005792	.008643
常数项+时间 [主题= probandn]	UN (1.1)	.005318	.002446	2.174	.030	.002159	.013099
	UN (2.1)	-.000733	.000776	-0.944	.345	-.002254	.000789
	UN (2.2)	.001603	.000429	3.734	.000	.000948	.002709

a. 因变量: Y。

表“协方差参数估计”反映了随机效应的估计值。由于该模型有两个参数，SPSS 对截距的方差（“UN（1.1）”）、斜率的方差（“UN（2.2）”）以及截距与斜率间的协方差（“UN（2.1）”）做出估计。所有方差参数达到了显著性水平（ $p=0.039$ 或 $P=0.000$ ），而协方差则未达到显著性水平（ $p=0.345$ ），即截距与斜率之间的协方差未达到显著性水平。因此，成绩水平与成绩斜率之间可能有的相互作用只会对少数个案，但不会对所有个案产生影响。由此可以认为，被试者在培训过程中以不同速度提升其成绩，此外个别被试者的成绩在不同水平上波动。这个调查结果通过将预测的测试成绩可视化得到证实，（不受未达到显著性水平的影响）足以使数据接受进一步分析。



这个图展示了三阶段分析的第三种结果：被试者具有不同的斜率和截距。这意味着向回推演到这项调查的设计方案：被试者的成绩起点不同，并且以不同的速度提高。例如，从这个图中可以看出，与培训开始时神经心理表现较差的被试者相比，神经心理表现较好的被试者在培训过程中的成绩提高幅度更大。所调查的个体生长曲线表明，39 名被试者可以分为两组，并做进一步分析，例如，通过重复测量对两组进行判别分析或方差分析。

5.3 岭回归 (SPSS 宏)

本章介绍了岭回归（参见 Chatterjee & Price, 1995², 228-240）。岭回归（主要）是一种直观的检验方法，即检验可能具有多重共线性的数据是否可以利用多元线性回归分析进行分析。

简而言之，即针对一个待计算初始模型计算出多个不同的模型变体，在这些模型变体中，初始的多重共线性程度（用 K 值，又称为岭迹表示）逐级下降。因此，K 值也就定义了 OLS 回归和岭回归之间渐进的差异。当 K=0 时，岭回归的估计值等于 OLS 回归的估计值（关于岭迹的推导请参见 Chatterjee& Price, 1995², 229, 236-240）。K 值在这里可以视为常数，分别添加到各预测变量方差中，从而降低了各个预测变量之间的相关性。

$$r_{12}^2 = \frac{[\sum x_1x_2 / (n-1)]^2}{s_{d_1}^2 s_{d_2}^2}$$

例如，如果将同样的常数分别添加到包括两个预测变量 x_n 及其方差 sd_n 的这个回归方程，则这两个预测变量之间的相关性和 VIF 值都会降低（参见 Cohen 等，2003³，428）。本章末尾归纳了选择适合 K 值的四条准则。

在呈现初始（未作改动的、最大的）的多重共线性时， $K=0$ 。随着 K 值的不断增大（如果从数据角度允许的话），多重共线性逐步降低。因此对于每个 K 值，根据某个初始模型计算出不同的模型变体，在每个模型变体中都调整了多重共线性的程度，从而也就根据多重共线性的程度调整了回归系数的估计值。如果从某个特定的 K 值开始，回归系数的估计值不再变化，或者只有细微变化，则可以认为具有稳定的回归估计，原则上可以像线性分析一样对这个回归估计进行解释。如果岭回归得出结论：这些数据不适用于线性回归分析，那么根据 Chatterjee & Price（1995²，229，222）的建议，作为代替可以使用很少变量或者主成分分析法（PCR），也可以考虑使用 PLS 回归。岭回归不能替代多元线性回归（Cohen 等人，2003³，427-428；Pedhazur，1982²，247）。

5.3.1 利用岭迹实现多重共线性的可视化

与其他统计方法相反，岭回归无法通过菜单导航，而是只能以 SPSS 宏“Ridge-Regression.sps”的形式提供。但是，岭回归的实施过程并不复杂。

```
GET
  FILE='C:\...\CP193.sav'.
INCLUDE "C:\...Ridge-Regression.sps" .

RIDGEREG DEP= Importe
/ENTER= InProd Lager Konsum
/START= 0
/STOP= 1
/INC= 0.05.
```

通过 GET FILE 命令调用 SPSS 数据集“CP193.sav”。利用全部的进口数据（ $N=18$ ）计算岭回归。这个数据集是以 Chatterjee & Price（1995²，193）著作中的进口数据为基础的。通过 INCLUDE 命令，将 SPSS 宏“Ridge-Regression”纳入分析。在 DEP=命令后，给定 IMPORTE 作为因变量；在 ENTER=命令后，给定 INPROD、LAGER 和 KONSUM 作为因变量。通过 START、STOP 和 INC 命令设定所输出岭迹（K 值）的极差和计数方式。在本例中，确定了对从 0 到 1 的 K 值应以 0.05 为一级逐步计数。在第 5.2.2 节，利用自选的 K 值计算出一个单独的岭回归。

需要指出的是，尽管这几个示例是建立在相同数据基础之上，但 Chatterjee & Price（1995²，193，232）和 SPSS 宏“Ridge-Regression”得到不同的结果（见下文）。该书出版时，尚未对这个现象的原因做出解释。

输出结果

表格“K 估计值的 R^2 和贝塔（B）系数”针对所调查的模型（潜在多重共线性预测变量 INPROD、LAGER 和 KONSUM 对于因变量 IMPORTE 的预测）分别反映了 K 值（岭迹）、

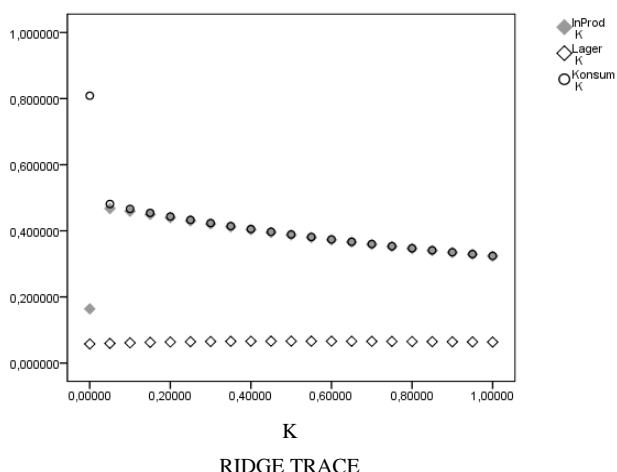
RSQ (R^2) 以及标准化回归系数贝塔 (β)。表格中每一行代表一次岭回归，分别通过逐步增大的 K 值予以定义。这样，表格“K 估计值的 R^2 和贝塔 (β) 系数”就反映了所有 21 次岭回归的贝塔 (β) 值和 K 值。由于通过这种方式很难将这些数值 (和模型) 相互比较，就用额外两个图将输出的模型参数可视化。图“RIDGE TRACE”反映了 K 值和贝塔 (β) 值，图“ R^2 与 K 值”反映了 K 值和 RSQ 值。

估计值的 R^2 和贝塔 (β) 集系

K	RSQ	InProd	Lager	Konsum
.00000	.97304	.163887	.057789	.808704
.05000	.97226	.467491	.059430	.481021
.10000	.97067	.459344	.061201	.466211
.15000	.96820	.449564	.062609	.454175
.20000	.96498	.439750	.063712	.443224
.25000	.96112	.430197	.064561	.432985
.30000	.95672	.420980	.065196	.423309
.35000	.95184	.412118	.065650	.414117
.40000	.94656	.403604	.065953	.405355
.45000	.94094	.395427	.066127	.396985
.50000	.93504	.387572	.066193	.388974
.55000	.92889	.380021	.066165	.381297
.60000	.92254	.372760	.066059	.373929
.65000	.91603	.365772	.065887	.366851
.70000	.90938	.359043	.065657	.360044
.75000	.90262	.352558	.065380	.353492
.80000	.89577	.346306	.065061	.347181
.85000	.88887	.340273	.064708	.341096
.90000	.88191	.334448	.064326	.335225
.95000	.87493	.328822	.063920	.329556
1.00000	.86793	.323382	.063493	.324080

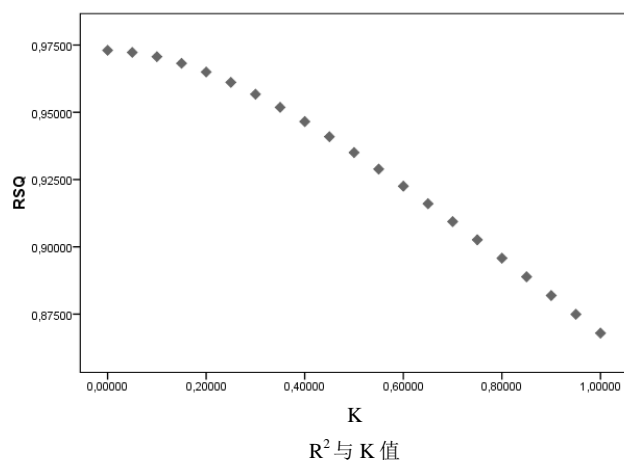
从表“K 估计值的 R^2 和贝塔 (β) 系数”可以看出，潜在多重共线性预测变量 INPROD、LAGER 和 KONSUM 的回归系数从 K=0.00 急剧变化到 K=0.05，在多重共线性不断减小的模型中仍只是做细微变化。从 K=0.90 开始，模型看起来变得稳定，INPROD 和 KONSUM 的回归系数在这之后只有小数点后两位的变化。预测变量 LAGER 看起来是总体最稳定的，因为从 K=0.05 开始就只有小数点后三位的变化。

图“RIDGE TRACE”反映了表格“K 估计值的 R^2 和贝塔 (β) 系数”中的 K 值和贝塔 (β) 值。分别由 X 轴表示相应预测变量 (例如，INPROD、LAGER 和 KONSUM) 的贝塔 (β) 值，由 Y 轴表示 K 值。



从图“RIDGE TRACE”可以看出，预测变量 INPROD 和 KONSUM 的回归系数从 $K=0.00$ 十分明显地变化到 $K=0.05$ 。之后 INPROD 和 KONSUM 的值相互重合。从 $K=0.90$ 开始，模型看起来变得稳定；估计回归系数的函数慢慢地转变为与 X 轴平行。

图“ R^2 与 K 值”反映了表格“ K 估计值的 R^2 和贝塔 (β) 系数”中的 K 值和 RSQ 值。由 X 轴表示相应预测变量（例如，INPROD、LAGER 和 KONSUM）的贝塔 (β) 值，由 Y 轴表示 RSQ 值。



从图“ R^2 与 K 值”可以看出，方差解释是否以及在多大程度上受到调节多重共线性的影响。当 $K=0.05$ 时， R^2 为 97%；而当 $K=0.9$ 时， R^2 下降到 88%（参见表格“ K 估计值的 R^2 和贝塔 (β) 系数”）。

Chatterjee & Price (1995², 233) 归纳了如何选择 K 值的四条准则。这些准则建立在 Hoerl & Kennard (1970) 论文基础之上，在这里表述如下。

1. 从某个 K 值开始，系统变得稳定。可以观察到：所估计回归系数的所有函数过渡到与 X 轴的方向平行。
2. 从内容上看，所测定的系数作为变化率应是可信的。

- 3. 从内容上看，所测定的系数应具有正确的正负号。
- 4. 模型参数（ R^2 ，残差）的质量不得下降到令人无法接受的程度。

下一节示范性地测定了所选择 K 值（例如， $K=0.05$ ）的模型参数。

5.3.2 岭回归的计算

代替作为示范选择的 K 值（例如， $K=0.05$ ），下面通过以下语句调用一次单独的岭回归，目的是演示：模型参数（ R^2 ，残差）如何急剧地从 $K=0.0$ （未画出）恶化到 $K=0.05$ 。岭回归不是多元线性回归的替代选择（Cohen 等人，2003³，427-428、Pedhazur，1982²，247）。这个例子只能演示 SPSS 编程的工作原理，不能未经改动就作为模板应用到其他统计分析。

语句：

```
RIDGEREG DEP= Importe
/ENTER= InProd Lager Konsum
/START= 0
/STOP= 1
/INC= 0.05
/K= 0.05.
```

在这里需要注意的是，在调用一次岭回归时，选项 START、STOP 和 INC 不起作用。

结果

Run MATRIX procedure:

***** Ridge-Regression with k = 0.05 *****

Mult R .986034205
RSquare .972263454
Adj RSqu .966319909
SE 2.290586357

ANOVA table			
	df	SS	MS
Regress	3.000	2574.856	858.285
Residual	14.000	73.455	5.247

F value	Sig F
163.5830760	.0000000

-----Variables in the Equation-----

	B	SE (B)	Beta	
B/SE (B)				
InProd	.09186403	.00588029	.46749148	15.62236235
Lager	.42595908	.31002801	.05942954	1.37393742

```
Konsum      .14438725      .00898625      .48102130  16.06757191
Constant -17.47526059      2.28784380      .00000000  -7.63831017
```

```
----- END MATRIX -----
```

在标题“Run MATRIX procedure (运行矩阵过程):”后面, 输出了 $K=0.05$ 时的一次岭回归的模型参数。“Mult R”、“RSquare” (0.9722, 等于表“K 估计值的 R^2 和贝塔 (B) 系数”中的数值)、“Adj Rsqu”和“SE”反映了多重 R、未调整 R^2 和调整 R^2 , 以及相应的估计误差。例如, “RSquare”只有从 0.0 ($R^2=0.973$) 到 0.05 ($R^2=0.972$) 的细微变化。 R^2 值的下降是可以接受的。

在标题“ANOVA table (方差分析表)”后面, 输出了方差分析的结果。这个表包含了回归或残差的平方和、均方和、F 值和 F 值的显著性。代入 $F=163.58$ 和 $p=0.000$, 该模型得到了统计显著性。例如, 残差平方和从 $K=0.0$ 模型 (模拟计算, 这里没有反映出) 时的 71.390 上升到 $K=0.5$ 模型时的 73.45。残差平方的增加是可以接受的。

在标题“方程中的变量”后面, 输出了 $K=0.05$ 时岭回归的结果。这个表包含了非标准化回归系数 (B)、标准化回归系数 (β) 和非标准化回归系数 B 的估计误差。所列出的标准化回归系数也就是表“估计值的 R^2 和贝塔 (B) 系数”中的 β 值。对此不再赘述。

5.3.3 SPSS 宏 “Ridge-Regression”

需要做的设置

```
GET FILE "C:\...\IHREDATEN.SAV" .
INCLUDE "C:\...\SPSS\Ridge-Regression.sps" .

RIDGEREG DEP= Y1
/ENTER = X1 X2 X3 .
/START= 0
/STOP= 1
/INC= 0.05
[/K= 0.05].
```

首先通过 GET FILE 命令调用一个数据集, 要纳入岭回归的变量就在这个数据集中。

通过 INCLUDE 命令将 SPSS 宏 “Ridge-Regression” 纳入分析程序。无须对宏进行拟合, 只需通过 INCLUDE 命令将其纳入即可。这个宏由 SPSS 提供, 多数情况下位于安装 SPSS 时自动创建的目录“C:\...\SPSS”中。

在 RIDGEREG 命令后, 只需给定因变量 (在 DEP=命令后), 在 ENTER=命令后只需给定自变量, 并针对这个自变量进行岭回归。通过 START、STOP 和 INC 命令设定所输出岭迹 (K 值) 的极差和计数方式。通过 K=命令, 能够调用一个具有某个 K 值的、单独的岭回归。在 K=命令之后, 只能给定一个数值。这里需要注意的是, 在调用一个岭回归时, 选项 START、STOP 和 INC 无效。更多信息请参见 SPSS 语法的技术文档。如果选项“分割文件” (SPLIT FILE) 起作用, 则不能执行宏 “Ridge-Regression”。由于疏忽, 2008 年的 SPSS 16 版没有提供宏 “Ridge-Regression”。

特征宏

```

preserve.
set printback=off.
define ridgereg (enter=!charend ('/'))
    /dep = !charend ('/'))
    /start=!default (0) !charend ('/'))
    /stop=!default (1) !charend ('/'))
    /inc=!default (.05) !charend ('/'))
    /k=!default (999) !charend ('/'))
    /debug=!DEFAULT ('N') !charend ('/')) ).

preserve.
!IF ( !DEBUG !EQ 'N') !THEN
set printback=off mprint off.
!ELSE
set printback on mprint on.
!IFEND .
SET mxloops=200.

*-----
* Save original active file to give back after macro is done.
*-----
!IF (!DEBUG !EQ 'N') !THEN
SET RESULTS ON.
DO IF $CASENUM=1.
PRINT / " NOTE: ALL OUTPUT INCLUDING ERROR MESSAGES HAVE BEEN TEMPORARILY "
      / " SUPPRESSED. IF YOU EXPERIENCE UNUSUAL BEHAVIOR, RERUN THIS "
      / " MACRO WITH AN ADDITIONAL ARGUMENT /DEBUG='Y'. "
      / " BEFORE DOING THIS YOU SHOULD RESTORE YOUR DATA FILE. "
      / " THIS WILL FACILITATE FURTHER DIAGNOSIS OF ANY PROBLEMS. " .
END IF.
!IFEND .

save outfile='rr__tmpl.sav'.

*-----
* Use CORRELATIONS to create the correlation matrix.
*-----

* DEFAULT: SET RESULTS AND ERRORS OFF TO SUPPRESS CORRELATION PIVOT TA-
BLE *.
!IF (!DEBUG='N') !THEN
set results off errors off.
!IFEND

correlations variables=!dep !enter /missing=listwise/matrix out (*) .
set errors on results listing .

```

```

*-----.
* Enter MATRIX.
*-----.

matrix.

*-----.
* Initialize k, increment, and number of iterations. If k was
* not specified, it is 999 and looping will occur. Otherwise,
* just the one value of k will be used for estimation.
*-----.

do if (!k=999) .
. compute k=!start.
. compute inc=!inc.
. compute iter=trunc ( (!stop - !start ) / !inc ) + 1.
. do if (iter <= 0) .
.   compute iter = 1.
. end if.
else.
. compute k=!k.
. compute inc=0.
. compute iter=1.
end if.

*-----.
* Get data from working matrix file.
*-----.

get x/file=*/names=varname/variable=!dep !enter.

*-----.
* Third row of matrix input is the vector of Ns. Use this to
* compute number of variables.
*-----.

compute n=x (3,1) .
compute nv=ncol (x) -1.

*-----.
* Get variable names.
*-----.

compute varname=varname (2: (nv+1) ) .

*-----.

```

```

* Get X'X matrix (or R, matrix of predictor correlations) from
* input data Also get X'Y, or correlations of predictors with
* dependent variable.
*-----.

compute xpx=x (5: (nv+4) ,2: (nv+1) ) .
compute xy=t (x (4,2: (nv+1) ) ) .

*-----.
* Initialize the keep matrix for saving results, and the names
* vector.
*-----.

compute keep=make (iter,nv+2,-999) .
compute varnam2={'K','RSQ',varname}.

*-----.
* Compute means and standard deviations. Means are in the
* first row of x and standard deviations are in the second
* row. Now that all of x has been appropriately stored,
* release x to maximize available memory.
*-----.

compute xmean=x (1,2: (nv+1) ) .
compute ybar=x (1,1) .
compute std=t (x (2,2: (nv+1) ) ) .
compute sy=x (2,1) .
release x.

*-----.
* Start loop over values of k, computing standardized
* regression coefficients and squared multiple correlations.
* Store results
*-----.

loop l=1 to iter.
. compute b = inv (xpx+ (k &* ident (nv,nv) ) ) *xy.
. compute rsq= 2* t (b) *xy - t (b) *xpx*b.
. compute keep (1,1) =k.
. compute keep (1,2) =rsq.
. compute keep (1,3: (nv+2) ) =t (b) .
. compute k=k+inc.
end loop.

*-----.
* If we are to print out estimation results, compute needed
* pieces and print out header and ANOVA table.

```

```

*-----
do if (!k <> 999) .
  !let !rrtitle=!concat ('***** Ridge-Regression with k = ',!k) .
  !let !rrtitle=!quote (!concat (!rrtitle,' ***** ')) .
  . compute sst=(n-1) * sy **2.
  . compute sse=sst * ( 1 - 2* t (b) *xy + t (b) *xpx*b) .
  . compute ssr = sst - sse.
  . compute s=sqrt ( sse / (n-nv-1) ) .
  . print /title=!rrtitle /space=newpage.
  . print {sqrt (rsq) ;rsq;rsq-nv* (1-rsq) / (n-nv-1) ;s}
  /rlabel='Mult R' 'RSquare' 'Adj RSquare' 'SE'
  /title=' '.
  . compute anova={nv,ssr,ssr/ (nv) ;n-nv-1,sse,sse/ (n-nv-1) }.
  . compute f=ssr/sse * (n-nv-1) / (nv) .
  . print anova
    /clabels='df' 'SS','MS'
    /rlabel='Regress' 'Residual'
    /title='          ANOVA table'
    /format=f9.3.
  . compute test=ssr/sse * (n-nv-1) /nv.
  . compute sigf=1 - fcdf (test,nv,n-nv-1) .
  . print {test,sigf} /clabels='F value' 'Sig F'/title=' '.

*-----
* Calculate raw coefficients from standardized ones, compute
* standard errors of coefficients, and an intercept term with
* standard error. Then print out similar to REGRESSION output.
*-----

. compute beta={b;0}.
. compute b= ( b &/ std ) * sy.
. compute intercpt=ybar-t (b) *t (xmean) .
. compute b={b;intercpt}.
. compute xpx= (sse/ (sst* (n-nv-1) ) ) *inv (xpx+ (k &*
                                ident (nv,nv) ) ) *xpx*
                                inv (xpx+ (k &* ident (nv,nv) ) ) .
. compute xpx= (sy*sy) * (mdiag (1 &/ std) *xpx*mdiag (1 &/ std) ) .
. compute seb=sqrt (diag (xpx) ) .
. compute seb0=sqrt ( (sse) / (n* (n-nv-1) ) + xmean*xpx*t (xmean) ) .
. compute seb={seb;seb0}.
. compute rnms={varname,'Constant'}.
. compute ratio=b &/ seb.
. compute bvec={b,seb,beta,ratio}.
. print bvec/title='-----Variables in the Equation-----'
  /rnames=rnms /clabels='B' 'SE (B)' 'Beta' 'B/SE (B)' .
. print /space=newpage.

```

```

end if.

*-----.
* Save kept results into file. The number of cases in the file
* will be equal to the number of values of k for which results
* were produced. This will be simply 1 if k was specified.
*-----.

save keep /outfile='rr__tmp2.sav' /names=varnam2.

*-----.
* Finished with MATRIX part of job.
*-----.

end matrix.

*-----.
* If doing ridge trace, get saved file and produce table and
* plots.
*-----.

!if (!k = 999) !then

get file='rr__tmp2.sav'.
print formats k rsq (f6.5) !enter (f8.6) .
report format=list automatic
  /vars=k rsq !enter
  /title=center 'R-SQUARE AND BETA COEFFICIENTS FOR ESTIMATED VALUES OF K'.

graph
  /title = 'RIDGE TRACE'
  /footnote = 'K'
  /scatterplot (overlay) k with !enter.

graph
  /title = 'R-SQUARE VS. K'
  /scatterplot k with rsq.

!ifend.

*-----.
* Get back original data set and restore original settings.
*-----.

get file=rr__tmp1.sav.
restore.
!enddefine.
restore.

```


第 6 章 其他方法和模型（一览）

SPSS 还为调用一个“回归”提供了许多其他方法。本章为此用一个简短的一览表，介绍了如何通过菜单或语句调用其他的回归方法。这个一览表使用户可以在选择和应用其他回归方法时把握大致的方向，对是否完整没有要求。本章开篇先介绍在 SPSS 菜单中的其他回归方法，最后再介绍这些方法在 SPSS 语句中的体现。本章的语句实例只起到解释和启发的作用，不能未经修改就用作其他分析的模板。建议在应用相应的回归方法之前，先大致了解一下回归方法（或者回归方法组合）的统计学特点。

这里介绍的其他用途仅限于 SPSS 16 版。其他版本的 SPSS 可能在菜单设计、功能范围和输出的语句（例如，使用 GENLOG 代替 LOGLINEAR）方面有所不同。

6.1 通过 SPSS 菜单调用其他回归方法

本节介绍了可以通过 SPSS 菜单调用的其他回归方法。左边列出的是 SPSS 16 版中的菜单。如果书中已经介绍了相应菜单中的某个回归方法，则在其右边有一条指出对应章节的提示。在此情况下既没有简述，也没有语句实例。如果通过某个菜单可以调用本书还没有介绍的、值得研究的回归方法，则右边会配有提示“参见下文”，然后通过简述和语句实例进行介绍。如果一个菜单无法调用回归方法，或者是在本书排印时这个菜单还不为人所知，则在回归方法、菜单和选项右边会提示“无说明”。本章没有列出菜单“生存...”，因为第 4 章已对其做了详细介绍。

菜单“回归”

菜单“回归”包含了一些其他回归方法（前提是取得所需 SPSS 模块的使用许可），例如，

“Probit”、“权重估计”、“两阶最小二乘法”和“最佳尺度”。



自动线性建模...	第 2.1、2.3 节
曲线估计	第 1.4.2、2.2.4 节
部分最小平方	第 5.1 节
二元 Logistic	第 3.1 节
多项 Logistic	第 3.3 节
有序	第 3.2 节
Probit	参见下文
非线性	第 2.2 节
权重估计	参见下文
两阶最小二乘法	参见下文
最佳尺度	参见下文

通过“Probit...”调用一个二元因变量的机率单位分析（通过“Logit”也可以）。机率单位分析差不多与逻辑回归一致，优先用于计划进行的实验，主要是为了在得分数据的基础上调查剂量-效应关系。相反，逻辑回归更适用于实证研究，主要用于胜率（Odds Ratio）的测定。

```
PROBIT
  n_responses OF n_obs WITH X1 X2
  /LOG NONE
  /MODEL BOTH .
```

“权重估计”。当线性回归方差齐次性的前提条件不满足时，人们会对使用权重估计法感兴趣。例如，线性回归的标准模型是以所调查的总体中呈现恒定方差（方差齐次性）为基础的。只要这个前提条件不能得到满足（如具有某种属性、数值较低的个案比数值较高的个案具有更大的方差时），那么使用普通最小二乘法（OLS 法），线性回归就只能得出次优的模型估计。如果变异性中的差异能够通过另一个变量预测出来（例如，通过 SOURCE 命令给定一个变量以及方差异次性的原因），就可以使用权重估计的加权最小二乘法（WLS 法）计算出一个线性回归模型的系数。

WLS 法就是通过将加权的残差平方和最小化来进行参数估计。而 OLS 法是通过将没有加权的残差平方和最小化进行参数估计。在使用 WLS 法时，根据与误差项方差倒数的比例选择加权。根据 Chatterjee & Price 的著作（1995²，53），对于经过转换的变量 y/x 和 $1/x$ 而言，使用 WLS 法和 OLS 法并无区别。

```
WLS
  Y1 WITH X1 X2
  /SOURCE HTSC
  /PRINT BEST.
```

“两阶最小二乘法”。当线性回归的某个前提条件不能满足时，人们同样会对使用两阶最小二乘法法（2SLS）感兴趣。例如，线性回归标准模型的前提是，标准中的误差和预测变量不相关。如果模型违反了这个前提（如变量之间产生交互作用），则 OLS 法同样只能得出次优

的模型估计。相反，2SLS 法使用的是与误差大小不相关的辅助变量（参照 STRUMENTS）。在第一阶段，测定有问题的预测变量的近似值，然后用这些近似值在第二阶段就可以计算出一个线性回归。计算结果最终是建立在与误差不相关的数值基础上的。

```
2SLS
  Y1 WITH X1, X2
/X1 WITH Y1, X3
/INSTRUMENTS=X2, X3.
```

“最佳尺度”。原则上是一个定量分类回归，在这个回归中通过 ALS 法（交替最小二乘法）为多个（定类、定序还有定距的）预测变量给定一个最佳的尺度水平。这样做能够为已转换变量计算出一个最佳的线性回归方程。这个方法的优点是：无须关于变量的分布性假设。

```
CATREG
VARIABLES=Y1 X1 X2
/ANALYSIS=Y1 (LEVEL=SPORD,DEGREE=2,INKNOT=2)
      WITH X1 (LEVEL=SPORD,DEGREE=2,INKNOT=2)
      X2 (LEVEL=SPORD,DEGREE=2,INKNOT=2)
/DISCRETIZATION=Y1 (GROUPING,NCAT=7,DISTR=NORMAL)
      X1 (GROUPING,NCAT=7,DISTR=NORMAL)
      X2 (GROUPING,NCAT=7,DISTR=NORMAL)
/MISSING=Y1 (LISTWISE) X1 (LISTWISE) X2 (LISTWISE)
/PRINT=RCOEFF ANOVA.
```

菜单“对数线性模型”

菜单“对数线性模型”包含了其他选项（前提是取得所需 SPSS 模块的使用许可）。例如，“常规”、“Logit”和“模型选择”。



常规 参见下文

Logit 参见下文

模型选择 参见下文

通过“**常规**”可以调用“一般对数线性”法（SPSS 过程命令 GENLOG）。这个一般的和多样化的方法是针对模型拟合、假设检验和对含有（自变量）因变量的交叉列表的频率（分类数据）进行参数估计而开发的。“一般对数线性”方法除了包含（非分层）分层的多维列联表外，还包含对分类变量的逻辑回归。

```
GENLOG Categ1 Categ2 WITH Covar
/MODEL= POISSON
/PRINT= FREQ RESID ADJRESID ZRESID DEV
/PLOT= RESID (ADJRESID) NORMPROB (ADJRESID)
/CRITERIA= CIN (95) ITERATE (20) CONVERGE (0.001) DELTA (.5)
/DESIGN Categ2 Covar*Categ2 Categ1 Covar*Categ1
/SAVE= ADJRESID.
```

通过“**Logit**”方法（“Logit-对数线性”，SPSS 过程命令 GENLOG）同样可以检验带有自变量和定类因变量的模型。因变量必须是定类的，而自变量可以是定类的，也可以不是。

“一般对数线性”法主要包含对分类变量的逻辑回归。旧版 SPSS 中的 LOGLINEAR 命令只有通过语句才能使用（如果要对 GENLOG、HILOGLINEAR 和 LOGLINEAR 做一个技术上和统计上的比较，可参见 Command Syntax Reference）。

```
GENLOG
  DosisGrd1 BY DosisGrd2
/MODEL= MULTINOMIAL
/PRINT= FREQ RESID ADJRESID ZRESID DEV
/PLOT= NONE
/CRITERIA= CIN (95) ITERATE (20) CONVERGE (0.001) DELTA (.5)
/DESIGN .
```

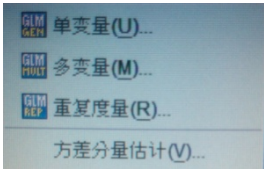
通过“模型选择”可调用“分层对数线性分析”方法（SPSS 过程命令 HILOGLINEAR）。通过这个方法可以分析多维表格（列联表），即通过用逐步法拟合分层对数线性模型来实现。

```
HILOGLINEAR
  DosisGrd1 (0 1) DosisGrd2 (0 1)
/CWEIGHT= DosisGrd1
/CRITERIA ITERATION (20) DELTA (.5)
/PRINT= FREQ RESID
/DESIGN .
```

对于分层模型来说，SPSS 过程命令 HILOGLINEAR 比诸如 GENLOG 等 SPSS 过程命令更为有效。但它无法生成不饱和模型的参数估计值，也不能对参数的对比进行定义。

“一般线性模型”

菜单“一般线性模型”包含（前提是取得所需 SPSS 模块的使用许可）如“单变量”、“多变量”、“重复测量”和“方差分量估计”等选项。



- 单变量 参见下文
- 多变量 参见下文
- 重复测量 无说明
- 方差分量估计 参见 MIXED

通过“单变量”，可以利用一个或多个因子和/或变量计算出一个因变量的回归分析。在这个过程中，因子变量将调查的总体分成各组。既可以检验平衡模型，也可以检验不平衡模型（参见菜单“方差分量估计”）。

```
GLM
  Y WITH X1 X2.
```

通过“多变量”可用一个或多个因子和/或变量计算出多个因变量的回归分析。在这个过程中，因子变量将调查的总体分成各组。既可以检验平衡模型，也可以检验不平衡模型。

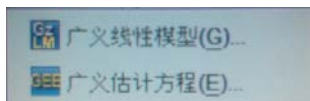
```
GLM
  Y1 Y2 WITH X1 X2 X3.
```

对于带有混合的模型（例如，在单变量的重复测量或带有随机组块的设计时），利用菜单

“方差分量估计”可以测定每个随机效应对于因变量方差的贡献率，从而可以在设计时进一步减小方差。“方差分量估计”（VARCOMP）本质上是 MIXED 的一个子集（参见下文）；因此两个菜单或者 SPSS 过程命令得出同样的方差估计值。但是，MIXED 的功能范围明显大于 VARCOMP，因此选择 MIXED。

菜单“广义线性模型”

菜单“广义线性模型”包含（前提是取得所需 SPSS 模块的使用许可）“广义线性模型”和“广义估计方程”等选项。



广义线性模型 见下文

广义估计方程 见下文

通过“广义线性模型”提供了一般线性模型的一种扩展形式。在这个模型中，通过一个待给定的链接函数来确定因子、协变量和因变量之间的关联，因变量也会具有不同于正态分布的分布。这样，广义线性模型就涵盖了常用的回归方法，例如，用于正态分布反应变量的线性回归、用于二元数据的逻辑回归模型、用于得分数据（如泊松回归，伽马回归）和截尾生存时间数据（如区间截尾的生存时间数据的互补重对数回归）的对数线性回归。广义线性模型通常不能用于相互关联的自变量。下面的例子展示了一个二元逻辑回归的 GENLIN 语句。

```
GENLIN
  Y1 BY X1
  /MODEL X1 DISTRIBUTION=BINOMIAL LINK=LOGIT
  /EMMEANS TABLES=X1 SCALE=ORIGINAL.
```

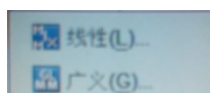
选项“广义估计方程”再次扩展了“广义线性模型”，使其可以包含相互关联的纵向数据。下面的例子展示了一个用于重复测量的逻辑回归的 GENLIN 语句。因此，与广义线性模型相比，“广义估计方程”选择菜单的主要区别在于：还可以对重复测量进行分析，例如，通过选项“重复”和“内在特性变量”。

```
GENLIN
  Y1 (REFERENCE=LAST)
  BY X1 X2 (ORDER=ASCENDING)
  /MODEL X1 X2 INTERCEPT=YES DISTRIBUTION=BINOMIAL
  LINK=LOGIT
  /CRITERIA METHOD=FISHER (1) CILEVEL=95
  /REPEATED SUBJECT=id WITHINSUBJECT=X2 SORT=YES CORR
  TYPE=UNSTRUCTURED ADJUSTCORR=YES COVB=ROBUST UPDATE
  CORR=1
  /MISSING CLASSMISSING=EXCLUDE
  /PRINT CPS DESCRIPTIVES MODELINFO FIT SUMMARY SOLUTION
  WORKINGCORR.
```

菜单“混合模型”

菜单“混合模型”目前有（前提是取得所需 SPSS 模块的使用许可）“线性...”和“广义”

两个选项。



线性... 见下文

广义... 见下文

只有当一个模型中出现混合因子，也就是当随机性因子和确定性因子同时出现时，才可以将这个模型称为“混合模型”（“mixed model”）或模型 III。混合模型（MIXED）可以用于具有收益的回归分析问题（参照“方差分量估计”下的提示，VARCOMP，见上文）。例如，与简单线性模型相反，混合模型能以随机截距模型的形式来描述个别个案。与不能根据描述个别个案特性的简单回归直线相反，随机截距模型为每一个个案生成一条独有的直线（即成绩特性曲线）。通常情况下，这种模型分为子模型“随机截距”、“随机斜率”和“随机截距与随机斜率”。通过 AIC 信息准则，能够相互直接比较这些模型。这个方法在第 5.2 节已做详细介绍。

示例：“随机截距”（假设：每个个案都有自己的截距）

```
MIXED
  Y1 WITH TIME
  /FIXED INTERCEPT TIME
  /RANDOM INTERCEPT | SUBJECT (ID) COVTYPE (ID)
  /PRINT SOLUTION TESTCOV.
```

示例：“随机斜率”（假设每个个案都有自己的斜率）。

```
MIXED
  Y1 WITH TIME
  /FIXED INTERCEPT TIME
  /RANDOM TIME | SUBJECT (ID) COVTYPE (ID)
  /PRINT SOLUTION TESTCOV.
```

示例：“随机截距与随机斜率”（假设斜率和截距是随机的）。

```
MIXED
  Y1 WITH TIME
  /FIXED INTERCEPT TIME
  /RANDOM INTERCEPT TIME | SUBJECT (ID) COVTYPE (UN)
  /PRINT SOLUTION TESTCOV.
```

SPSS 过程命令 MIXED 性能很强，除了裂区设计外，还可以进行多层回归分析。

菜单“时间序列”

对于具有时间相依性或者季节性结构和预测的时间序列数据和纵向数据，可使用 OLS 回归（例如，参见 Woolridge 的著作，2003，第 10 和 11 章；Cohen 等人的著作，2003³，第 15 章；Chatterjee 和 Price 的著作，1995²，第 7 章）。此外，也可以使用专门的时间序列分析方法（例如参见 Hartung 的著作，1999，XII；Schlittgen，2001；Schlittgen & Streitberg，2001⁹；Yaffee & McGee，2000）。菜单“时间序列”包含（前提是取得所需 SPSS 模块的使用许可）

“创建模型”、“自回归（AR）”和“ARIMA”等选项。



创建模型	参见下文
应用模型	无说明
季节性分解	无说明
频谱分析	无说明
序列图	无说明
自相关	无说明
互相关	无说明

菜单“创建模型”包含“Expert Modeler”，它针对一个或多个相依时间序列，自动测定和估计出一个最佳拟合的 ARIMA 模型或指数平滑模型。因此，用户无须费力地试着建立最佳模型。用户也可将自定义的 ARIMA 模型或指数平滑模型传递给 SPSS。作为拟合优度的标准，主要是输出 R^2 标准、RMSE、MAE、MAPE、MaxAE、MaxAPE 和 BIC（Bayes 信息准则）。

```
DATE DAY 1 3.
PREDICT THRU CYCLE 5 DAY 3 .

TSMODEL
/MODELSUMMARY
    PRINT=[MODELFIT RESIDACF RESIDPACF]
    PLOT=[SRSQUARE RSQUARE RMSE MAPE MAE]
/MODELSTATISTICS DISPLAY=YES
    MODELFIT=[SRSQUARE RSQUARE RMSE MAPE MAE]
/MODELDETAILS
    PRINT=[PARAMETERS RESIDACF RESIDPACF FORECASTS]
    PLOT= [RESIDACF RESIDPACF]
/SERIESPLOT OBSERVED FORECAST FIT FORECASTCI FITCI
/OUTPUTFILTER DISPLAY=ALLMODELS
/AUXILIARY CILEVEL=95 MAXACFLAGS=24
/SAVE PREDICTED (Vorhersagewert)
/MISSING USERMISSING=EXCLUDE
/MODEL DEPENDENT=Y1 Y2 Y3 PREFIX='Skala'
/EXPERTMODELER TYPE=[ARIMA EXSMOOTH] TRYSEASONAL=YES
/AUTOOUTLIER DETECT=ON TYPE=[ ADDITIVE LEVELSHIFT].
```

“自回归”提供了多个利用一阶自回归误差估计线性回归模型的方法。SPSS 为此主要提供了布朗指数平滑法和温特指数平滑法。

```
TSMODEL
/MODEL DEPENDENT=Y1 Y2 INDEPENDENT=X1
/EXSMOOTH TYPE=WINTERSADDITIVE.
```

SPSS 中的“ARIMA”（自回归移动平均）提供了很多方法来估计非季节性和季节性单变量模型（即 Box-Jenkins 模型）。可以转换数据序列，也可以设置周期性和其他的季节性参数。

```
TSMODEL
/AUXILIARY SEASONLENGTH=12
```

```

/MODEL DEPENDENT=Y1
/ARIMA AR=[0] ARSEASONAL=[0] MA=[1] MASEASONAL=[1]
DIFF=1 DIFFSEASONAL=1 TRANSFORM=LN.

```

Expert Modeler 会自动测定一个或多个时间序列的最优拟合模型。自 SPSS 第 14 版起, Expert Modeler (SPSS 过程命令 TSMODEL) 替代了 SPSS 过程命令 AREG 和 ARIMA。用户无须再费力地“亲手”摸索最适合的模型。TSMODEL 的选项 EXSMOOTH 主要可以拟合带有 AR(1) 误差(自回归一阶误差)的回归模型。相反, 通过 ARIMA 选项可以拟合带有或者没有确定性回归量的季节性(非季节性)单变量 ARIMA 模型。

6.2 可用语句调用的其他回归形式

本节介绍了可用 SPSS 语句调用的其他回归形式, 本章也介绍了直接通过 SPSS 菜单调用的回归方法。语句实例帮助人们理解那些通常通过编程, 尤其是通过 SPSS 命令语法参考实现的、其他被低估的 SPSS 功能。文中例子仅起到解释说明的作用, 不能未经改动就作为统计分析的模板。

通过 SPSS 过程命令 ANOVA, 可以用一个或几个因子和/或变量计算出一个因变量的回归分析。

```

ANOVA
VARIABLES= Y1 BY X1 (0,1) WITH X2
/METHOD= UNIQUE
/STATISTICS= REG.

```

通过 SPSS 过程命令 MANOVA, 可用一个或多个因子和/或多个变量计算出多个因变量的多元线性回归分析。

```

MANOVA
Y1 Y2
WITH X1 X2 X3.

```

通过 SPSS 过程命令 UNIANOVA, 可用一个或多个因子和/或变量计算出一个因变量的回归分析。

```

UNIANOVA
Y1 BY X1 WITH X2 .

```

通过 SPSS 过程命令 LOGLINEAR, 可调用对分类变量的逻辑回归(目前只能通过语句进行)。

```

LOGLINEAR
DosisGrd1 (1,3) DosisGrd2 (1,8)
/DESIGN= DosisGrd1, DosisGrd2.

```

LOGLINEAR 假设呈现多项分布。当表不完整时, LOGLINEAR 的重新参数化方法可能会得出错误的自由度, 从而得出错误的分析结果。

附录 A 公式

为了评估 SPSS 的输出结果，了解其统计定义和推导过程是必不可少的。因此为了完整起见，下面归纳了本书所涉及的主要统计方法的公式。这个一览表给出了相关计算方法和相应的定义，与 SPSS 的《SPSS 算法》手册以及《命令语法参考》手册（例如， SPSS，2007a，c）一致。感兴趣的读者可自行查阅。

相关分析的公式

SPSS 过程 CORRELATIONS

计算方法

N	案例数量
X_{kl}	案例 l 的变量 k 的值
W_k	纳入变量 k 统计量计算的案例的加权和
W_{kj}	纳入变量 k 与 j 统计量计算的案例的加权和
w_l	案例 l 的权重
平均值和标准差	$\overline{X}_k = \frac{\sum_{l=1}^N w_l X_{kl}}{W_k},$ $S_k = \sqrt{\left(\sum_{l=1}^N w_l X_{kl}^2 - \overline{X}_k^2 W_k\right) / (W_k - 1)}$
叉积偏差和协方差	$C_{ij} = \sum_{l=1}^N w_l X_{il} X_{jl} - \left(\sum_{l=1}^N w_l X_{il}\right) \left(\sum_{l=1}^N w_l X_{jl}\right) / W_{ij}$
Pearson 相关	$r_{ij} = \frac{C_{ij}}{\sqrt{C_{ii} C_{jj}}}$

Cox 回归公式

SPSS 过程 COXREG

计算方法

风险函数

写法及变体

例如，一个在 j 层的个体的风险函数定义为：

生存函数（风险率的比例已给定）：写法及变体

β 值估计

偏似然函数中的对数似然

l 的一阶导数

基线函数的估计

一个单独变量的 R（如果 Wald > 2，否则，R 设定为零）
具有多个类别的一个变量的 R（如果 Wald > 2df）

曲线拟合公式

SPSS 过程 CURVEFIT

计算方法

Y_t

$E(Y_t)$

\hat{Y}_t

观察值序列 $t = 1, \dots, n$
 Y_t 的期望值
 Y_t 的预测值

$$\begin{aligned} h(t|\mathbf{x}) &= h_0(t)e^{\mathbf{x}'\beta} \\ h_j(t|\mathbf{x}) &= h_{0j}(t)e^{\mathbf{x}'\beta} \\ h(t|\mathbf{x}) &= \frac{f(t|\mathbf{x})}{S(t|\mathbf{x})} \\ S(t|\mathbf{x}) &= \int_t^\infty f(u|\mathbf{x})du \\ S(t|\mathbf{x}) &= [S_0(t)]^{\exp(\mathbf{x}'\beta)} \\ S_0(t) &= \exp(-H_0(t)) \\ l = \ln L(\beta) &= \sum_{j=1}^m \sum_{i=1}^{k_j} \mathbf{s}'_{ji}\beta - \sum_{j=1}^m \sum_{i=1}^{k_j} d_{ji} \ln \left(\sum_{l \in R_{ji}} w_l e^{\mathbf{x}'_l \beta} \right) \\ D_{\beta_r} = \frac{\partial l}{\partial \beta_r} &= \sum_{j=1}^m \sum_{i=1}^{k_j} \left(S_{ji}^{(r)} - d_{ji} \frac{\sum_{l \in R_{ji}} w_l x_{lr} e^{\mathbf{x}'_l \beta}}{\sum_{l \in R_{ji}} w_l e^{\mathbf{x}'_l \beta}} \right), \quad r = 1, \dots, p \\ \sum_{l \in D_i} \frac{w_l \exp(\mathbf{x}'_l \beta)}{1 - \alpha_i} &= \sum_{l \in R_i} w_l \exp(\mathbf{x}'_l \beta) \quad i = 1, \dots, k \\ R &= \left[\frac{\text{Wald-2}}{-2 \log\text{-likelihood for the initial model}} \right]^{1/2} \times \text{sign of MPLE} \\ R &= \left[\frac{\text{Wald-2*df}}{-2 \log\text{-likelihood for the initial model}} \right]^{1/2} \end{aligned}$$

模型	线性方程	因变量	自变量	系数
线性	$E(Y_t) = \beta_0 + \beta_1 t$	Y	t	β_0, β_1
对数	$E(Y_t) = \beta_0 + \beta_1 \ln(t)$	Y	$\ln(t)$	β_0, β_1

续表

模型	线性方程	因变量	自变量	系数
逆	$E(Y_t) = \beta_0 + \beta_1/t$	Y	$1/t$	β_0, β_1
平方	$E(Y_t) = \beta_0 + \beta_1 t + \beta_2 t^2$	Y	t, t^2	$\beta_0, \beta_1, \beta_2$
立方	$E(Y_t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$	Y	t, t^2, t^3	$\beta_0, \beta_1, \beta_2, \beta_3$
复合	$E(Y_t) = \beta_0 \beta_1^t$	$\ln(Y)$	t	β_0^*, β_1^*
幂	$E(Y_t) = \beta_0 t^{\beta_1}$	$\ln(Y)$	$\ln(t)$	β_0^*, β_1
S	$E(Y_t) = \exp(\beta_0 + \beta_1/t)$	$\ln(Y)$	$1/t$	β_0, β_1
生长	$E(Y_t) = \exp(\beta_0 + \beta_1 t)$	$\ln(Y)$	t	β_0, β_1
指数	$E(Y_t) = \beta_0 e^{\beta_1 t}$	$\ln(Y)$	t	β_0^*, β_1
逻辑	$E(Y_t) = (\frac{1}{u} + \beta_0 \beta_1^t)^{-1}$	$\ln(1/y - 1/u)$	t	β_0^*, β_1^*

Kaplan-Meier 法的公式

SPSS 过程 KM

计算方法

生存分布估计

$$\hat{S}(t) = \prod_{t_i < t} \left(1 - \frac{d_i}{n_i}\right)$$

平均生存时间估计

$$\hat{\mu} = \begin{cases} \sum_{i=0}^{k-1} \hat{S}(t_i^+)(t_{i+1} - t_i) \\ \sum_{i=0}^{k-1} \hat{S}(t_i^+)(t_{i+1} - t_i) + \hat{S}(t_k^+)(T_L - t_k) \end{cases}$$

逻辑回归公式

SPSS 过程 LOGISTIC REGRESSION

计算方法

n	观察案例的数量
P	参数数量
Y	含有元素 y_i (即二分因变量的案例 i 的观察值) 的 $n \times 1$ 向量
X	含有元素 xy_{ij} (即参数 i 的观察值) 的 $n \times p$ 矩阵
β	含有元素 β_j (即参数 j 的系数) 的 $p \times 1$ 向量
w	含有元素 w_i (即案例 i 的权重) 的 $n \times 1$ 向量
l	似然函数
L	对数似然函数
I	信息矩阵

似然函数	$l = \prod_{i=1}^n \pi_i^{w_i y_i} (1 - \pi_i)^{w_i (1 - y_i)}$
对数似然函数	$L = \ln(l) = \sum_{i=1}^n (w_i y_i \ln(\pi_i) + w_i (1 - y_i) \ln(1 - \pi_i))$
ML 估计值	$\sum_{i=1}^n w_i (y_i - \hat{\pi}_i) x_{ij} = 0$
-2 对数似然（在采用逐步法时）	$-2 \sum_{i=1}^n (w_i y_i \ln(\hat{\pi}_i) + w_i (1 - y_i) \ln(1 - \hat{\pi}_i))$
拟合优度	$\sum_{i=1}^n \frac{w_i (y_i - \hat{\pi}_i)^2}{\hat{\pi}_i (1 - \hat{\pi}_i)}$
Cox & Snell R ²	$R_{CS}^2 = 1 - \left(\frac{l(0)}{l(\hat{\beta})} \right)^{\frac{2}{W}}$
Nagelkerke R ² ，且满足 $\max(R_{CS}^2) = 1 - \{l(0)\}^{2/W}$	$R_N^2 = R_{CS}^2 / \max(R_{CS}^2)$
Hosmer-Lemeshow 检验	$\chi_{HL}^2 = \sum_{k=1}^g \frac{(O_{1k} - E_{1k})^2}{E_{1k} (1 - \xi_k)}$
偏 R（方程中变量的信息基础， 如果预测变量不是定类的）	$Partial R = \begin{cases} sign(\hat{\beta}_i) \sqrt{\frac{Wald_i - 2}{-2L(initial)}} \\ 0 \end{cases}$
偏 R（如果预测变量是定类的）	$Partial R = \begin{cases} \sqrt{\frac{Wald_i - 2(m-1)}{-2L(initial)}} \\ 0 \end{cases}$
多元回归公式	
SPSS 过程 NOMREG	
计算方法	
Y	具有整数值 1 到 J 的反应变量
J	名义反应类别的数量
M	子总体的数量
X	含有向量元素 x _i （即第 i 个子总体的位置模型的自变量观察值）的 m×p 矩阵
f _{ij} s	在子总体 i 上，Y=j 时属于单元格的第 s 个观察值的频率权重
n _{ij}	在子总体 i 上，Y=j 时属于单元格的观察值的频率加权
N	所有 n _{ij} s 的总和
π _{ij}	在子总体 i 上，Y=j 时的单元格概率
log (π _{ij} / π _{ik})	以反应类别 k 为底的反应类别 j 的对数
β _j = (β _{j1} , ..., β _{jp})	第 j 个对数中未知参数的 p×1 向量（例如，以反应类别 J 为底的反应类别 j 的对数）

$$\mathbf{B} = (\beta'_1, \dots, \beta'_{J-1})'$$

未知参数的 $(J-1) p \times 1$ 向量

$$\hat{\mathbf{B}} = (\hat{\beta}'_1, \dots, \hat{\beta}'_{J-1})'$$

B 的似然估计值最大值

$$\hat{\pi}_{ij}$$

π_{ij} 的似然估计值最大值

广义对数模型:

写法与变体

$$\pi_{ij} = \frac{\exp(\mathbf{x}'_i \beta_j)}{1 + \sum_{k=1}^{J-1} \exp(\mathbf{x}'_i \beta_k)}$$

$$\log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = \mathbf{x}'_i \beta_j, \text{ 对于 } j=1, \dots, J-1$$

模型对数似然

$$\begin{aligned} l(\mathbf{B}) &= \sum_{i=1}^m \sum_{j=1}^J n_{ij} \log(\pi_{ij}) \\ &= \sum_{i=1}^m \sum_{j=1}^J n_{ij} \log\left(\frac{\exp(\mathbf{x}'_i \beta_j)}{1 + \sum_{k=1}^{J-1} \exp(\mathbf{x}'_i \beta_k)}\right) \end{aligned}$$

最终模型: -2 对数似然

$$-2l(\tilde{\pi}) = -2 \sum_{i=1}^m \sum_{j=1}^J n_{ij} \log(\hat{\pi}_{ij})$$

模型的卡方

$$-2l(\tilde{\pi}) - \{-2l(\hat{\pi})\}$$

Cox & Snell R^2

$$R^2_{CS} = 1 - \left(\frac{L(\tilde{\pi})}{L(\hat{\pi})}\right)^{\frac{2}{n}}$$

Nagelkerke R^2

$$R^2_N = \frac{R^2_{CS}}{1 - L(\tilde{\pi})^{2/n}}$$

McFadden R^2

$$R^2_M = 1 - \left(\frac{l(\tilde{\pi})}{l(\hat{\pi})}\right)$$

皮尔逊拟合优度

$$X^2 = \sum_{i=1}^m \sum_{j=1}^J \frac{(n_{ij} - n_i \hat{\pi}_{ij})^2}{n_i \hat{\pi}_{ij}}$$

偏差拟合优度

$$D = 2 \sum_{i=1}^m \sum_{j=1}^J n_{ij} \log\left(\frac{n_{ij}}{n_i \hat{\pi}_{ij}}\right)$$

Wald 统计量 (包括置信区间)

$$\text{Wald}_{js} = \frac{\hat{B}_{js}}{\hat{\sigma}_{js}}, \quad \hat{B}_{js} \pm z_{1-\alpha/2} \hat{\sigma}_{js}$$

部分最小平方 (PLS) 回归公式 SPSS 过程 PLS

计算方法

X

自变量的 $N \times n$ 设计矩阵, 已中心化并且可能已标准化。
注意: 无截距

Y

因变量的 $N \times m$ 设计矩阵, 已中心化并且可能已标准化。

c

权重的 $m \times 1$ 列向量

u

Y 得分的 $N \times 1$ 列向量

w

权重的 $n \times 1$ 列向量

t

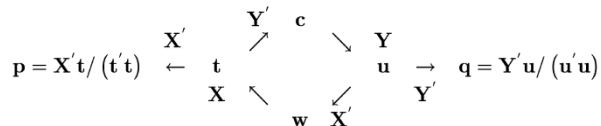
X 得分的 $N \times 1$ 列向量

d	待提取的 PLS 因子数量
p	$n \times 1$ 载荷向量
q	$m \times 1$ 载荷向量
P	$n \times d$ 载荷矩阵
Q	$m \times d$ 载荷矩阵
T	$N \times d$ 得分矩阵, $T = XW^*$
U	$N \times d$ 得分矩阵
W	X 权重的 $n \times d$ -矩阵
W^*	以原始坐标表示的 X 权重的 $n \times d$ 矩阵。这些权重可直接用于 X , $W^* W = (P'W)^{-1}$
C	Y 权重的 $m \times d$ 矩阵。这些权重可直接用于 Y 。
B	回归参数的 $n \times m$ 矩阵, $B = W^*C'$
E	残差的 $N \times n$ 矩阵, $E = X - TP'$
F	残差的 $N \times m$ 矩阵, $F = Y - UQ' = Y - XB$
DModX	X 变量与模型距离的 $N \times 1$ 向量
DModY	Y 变量与模型距离的 $N \times 1$ 向量
VIP	VIP (变量投影重要性) 的 $n \times d$ 矩阵

PLS 算法

在只有一个因变量 Y ($m = 1$ 时) 时, 使用 NIPALS 算法, 只需要一次迭代。在超过一个因变量 ($M > 1$) 时, 解出对等的特征问题。下面仅给出为 NIPALS (非线性迭代部分最小平方) 选取的公式。

右图展示了在 NIPALS 算法中向量和矩阵之间的关系。



可以在任何合理的步骤进入右侧所给出的循环。尤其是当 $m = 1$ 和 $c = 1$ 时, 可以进入具有 $u = Y$ 的步骤 1。循环重复执行, 直到到达最终收敛。

1. $w = X'u / (u'u)$
2. $w := w / \|w\|$
3. $t = Xw$
4. $c = Y't / (t't)$
5. $c := c / \|c\|$
6. $u = Yc$

尽管 NIPALS 算法在实际运用中被特征问题的解所取代, 但是通过图中定义的关系可以推导出所有必需的矩阵和向量。

T 对于 X 的回归和 Y 对于 u 的回归:

$$p = X't / (t't)$$

$$q = Y'u / (u'u)$$

X 矩阵和 Y 矩阵的减少:

$$X := X - tp'$$

$$Y := Y - tc' \quad (c \text{ 基于第四步, 非第五步, 见上文})$$

用于从 X 预测 Y 的回归系数的矩阵:

$$\begin{aligned} \mathbf{B} &= \mathbf{W}^* \mathbf{C}' \\ \mathbf{B} &= \mathbf{W} (\mathbf{P}' \mathbf{W})^{-1} \mathbf{C}' \\ \mathbf{B} &= \mathbf{X}' \mathbf{U} (\mathbf{T}' \mathbf{X} \mathbf{X}' \mathbf{U})^{-1} \mathbf{T}' \mathbf{Y} \end{aligned}$$

矩阵由这三个方程中的一个给定, 与 \mathbf{T} 和 \mathbf{U} 的尺度无关。

PLS 回归方程的解

占 Y 方差的比例, 该方差通过提取系数 k 得到解释:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\mathbf{B} + \mathbf{F} \\ SS_k(\mathbf{Y}) &= (\mathbf{t}'_{(k)} \mathbf{t}_{(k)}) \cdot \text{trace}(\mathbf{c}_{(k)} \mathbf{c}'_{(k)}) \\ &= (\mathbf{t}'_{(k)} \mathbf{t}_{(k)}) \cdot (\mathbf{c}'_{(k)} \mathbf{c}_{(k)}) \end{aligned}$$

$$VarProp_k(\mathbf{Y}) = \frac{SS_k(\mathbf{Y})}{\text{trace}(\mathbf{Y}' \mathbf{Y})}$$

已解释 Y 方差的累积比例:

$$CumVarProp_k(\mathbf{Y}) = \sum_{i=1}^k Var_i(\mathbf{Y})$$

占 X 方差的比例, 该方差通过提取系数 k 得到解释:

$$\begin{aligned} SS_k(\mathbf{X}) &= (\mathbf{t}'_{(k)} \mathbf{t}_{(k)}) \cdot \text{trace}(\mathbf{p}_{(k)} \mathbf{p}'_{(k)}) \\ &= (\mathbf{t}'_{(k)} \mathbf{t}_{(k)}) \cdot (\mathbf{p}'_{(k)} \mathbf{p}_{(k)}) \end{aligned}$$

$$VarProp_k(\mathbf{X}) = \frac{SS_k(\mathbf{X})}{\text{trace}(\mathbf{X}' \mathbf{X})}$$

已解释 X 方差的累积比例:

$$CumVarProp_k(\mathbf{X}) = \sum_{i=1}^k Var_i(\mathbf{X})$$

VIP (变量投影重要性) 统计量

对于各个变量和潜在因子计算出 VIP (变量投影重要性) 统计量, 作为其与模型的距离。

$$VIP_{jk} = \sqrt{\frac{n \sum_{l=1}^k w_{jl}^{*2} \cdot SS_l(\mathbf{Y})}{\sum_{l=1}^k SS_l(\mathbf{Y})}}$$

与模型的距离 (又称: DmodX, DmodY):

$$\begin{aligned} DModX_i &= \sqrt{\mathbf{e}'_i \mathbf{e}_i} \\ DModY_i &= \sqrt{\mathbf{f}_i' \mathbf{f}_i} \end{aligned}$$

PRESS 统计量 (例如)

$$PRESS = \sum_{i=1}^N DModY_i^2$$

有序回归公式

SPSS 过程 PLUM

计算方法

Y

假设整数值 1 到 J 的反应变量

J

定序反应的类别数量

M

子总体的数量

X^A

含有向量元素 x_i^A (即在第 i 个子总体的观察值, 由 SPSS 过程命令中的自变量来确定) 的 $m \times p^A$ 矩阵

X

含有向量元素 x_i (即在第 i 个子总体的位置模型自变量的观察值) 的 $m \times p$ 矩阵

Z

含有向量元素 x_i (即在第 i 个子总体的尺度模型自变量的观

f_{ijs}

n_{ij}

r_{ij}

n_i

N

π_{ij}

γ_{ij}

θ

β

τ

$\mathbf{B}=(\theta^T, \beta^T, \tau^T)^T$

$\hat{\mathbf{B}}=\left(\hat{\theta}^T, \hat{\beta}^T, \hat{\tau}^T\right)^T$

$\check{\mathbf{B}}=\left(\check{\theta}^T, \check{\beta}^T\right)^T$

一般模型

可能出现的链接函数:

对数、互补重对数、负重对数、Probit、Cauchit (逆柯西)。

对数似然函数

模型的对数似然

与

以及

模型信息

最终模型的-2 对数似然

模型的卡方统计量 (例如, 一般模型 vs 截距模型)

Cox & Snell R^2

察值) 的 $m \times p$ 矩阵

在子总体 i 中, $Y=j$ 时属于单元格的第 s 个观察值的频率权重

在子总体 i 中, $Y=j$ 时属于单元格的观察值的频率加权

在子总体 i 上, 直到并且包括 $Y=j$ 时的累积总量

子总体 i 的边界频率

所有频率的加权和

子总体 i 上, $Y=j$ 时的单元格概率

子总体 i 上, 直到并且包括 $Y=j$ 时的累积反应概率

在模型的位置部分中, 阈值参数的 $(J-1) \times 1$ 向量

在模型的位置部分中, 位置参数的 $p \times 1$ 向量

在模型的尺度部分中, 尺度参数的 $q \times 1$ 向量

一般模型中未知参数的 $\{(J-1)+p+q\} \times 1$ 向量

一般模型中参数 ML 估计值的 $\{(J-1)+p+q\} \times 1$ 向量

纯位置模型中参数 ML 估计值的 $\{(J-1)+p\} \times 1$ 向量

$$\eta_{ij}=\frac{\theta_j-\mathbf{b}^T \mathbf{x}_i}{\sigma\left(z_i\right)}, \quad \eta_{ij}=\operatorname{link}\left(\gamma_{ij}\right)$$

$$\operatorname{link}(\gamma)=\left\{\begin{array}{l} \log \left(\frac{\gamma}{1-\gamma}\right) \\ \log (-\log (1-\gamma)) \\ -\log (-\log (\gamma)) \\ \Phi^{-1}(\gamma) \\ \tan (\pi(\gamma-0.5)) \end{array}\right.$$

$$l=\sum_{i=1}^m \sum_{j=1}^{J-1} r_{ij} \varphi_{ij}-r_{i(J+1)} g\left(\varphi_{ij}\right), \quad \text { wobei } r_{ij}=\sum_{k=1}^j n k$$

$$\varphi_{ij}=\log \left(\frac{\gamma_{ij}}{\gamma_{ij+1}-\gamma_{ij}}\right),$$

$$g(\varphi)=\log \left(1+\exp (\varphi)\right)=\log \left(\frac{\gamma_{ij+1}}{\gamma_{ij+1}-\gamma_{ij}}\right).$$

$$-2 l\left(\hat{\mathbf{B}}\right)$$

$$-2 l\left(\mathbf{B}^{(0)}\right)-2 l\left(\hat{\mathbf{B}}\right)$$

$$R_{\text {CS }}^2=1-\left(\frac{L\left(\mathbf{B}^{(0)}\right)}{L\left(\hat{\mathbf{B}}\right)}\right)^{\frac{2}{n}}$$

Nagelkerke R^2

$$R_N^2 = \frac{R_{CS}^2}{1 - L(\mathbf{B}^{(0)})^{2/n}}$$

McFadden R^2

$$R_M^2 = 1 - \left(\frac{l(\hat{\mathbf{B}})}{l(\mathbf{B}^{(0)})} \right)$$

皮尔逊拟合优度

$$X^2 = \sum_{i=1}^m \sum_{j=1}^J \frac{(n_{ij} - n_i \hat{\pi}_{ij})^2}{n_i \hat{\pi}_{ij}}$$

偏差作为拟合优度

$$D = 2 \sum_{i=1}^m \sum_{j=1}^J n_{ij} \log \left(\frac{n_{ij}}{n_i \hat{\pi}_{ij}} \right)$$

参数统计量

Wald 统计量

$$\text{Wald}_k = \frac{\hat{B}_k}{\hat{\sigma}_k}$$

线性假设检验

$$\text{Wald}(\mathbf{L}, \mathbf{c}) = (\mathbf{L}\hat{\mathbf{B}} - \mathbf{c})^T \{ \mathbf{L} \text{Cov}(\hat{\mathbf{B}}) \mathbf{L}^T \}^{-1} (\mathbf{L}\hat{\mathbf{B}} - \mathbf{c})$$

线性回归公式

SPSS 过程 REGRESSION

计算方法

 y_i 具有方差 σ^2 / g_i 的案例 i 的因变量 c_i 案例 i 的权重, 如果通过 CASEWEIGHT 做了无偏差设定, 则 $c_i = 1$ g_i 案例 i 的回归权重, 如果通过 REGWGT 做了无偏差设定, 则 $g_i = 1$ L

不同案例的数量

 w_i $c_i g_i$ W

$$\sum_{i=1}^l w_i$$
, 所有案例的加权和
 P

自变量的数量

 C

$$\sum_{i=1}^l c_i$$
, 案例的加权和
 x_{ki} 案例 i 的第 k 个自变量 \bar{X}_k 第 k 个自变量的样本平均值:

$$\bar{X}_k = \left(\sum_{i=1}^l w_i x_{ki} \right) / W$$

 \bar{Y}

因变量的样本平均值:

$$\bar{Y} = \left(\sum_{i=1}^l w_i y_i \right) / W$$

 h_i 案例 i 的杠杆作用

\tilde{h}_i S_{ki} S_{yy} S_{ky} p^*

R

描述性统计量

R

多个 R

 R^2 调整 R^2 R^2 变化 (如果添加或移除了一个自变量块 q) F 值变化和相应的显著性统计量 (上面公式: 添加 q 个自变量; 下面公式: 移除 q 个自变量)。

残差平方和

回归平方和

赤池信息准则 (AIC)

施瓦兹信息准则 (BIC)

VIF 值 (VIF: 方差膨胀因子)

允差值

方程中变量的回归系数

系数的 95% 置信区间

截距 (用于含有截距模型):

 β 系数 $\frac{g_i}{W} + h_i$, 案例 i 杠杆作用的估计值 X_k 和 X_j 的样本协方差 Y 的样本方差 X_k 和 Y 样本协方差模型中的参数数量。若方程中不含截距, 则 $p^* = p$, 否则 $p^* = p + 1$ X_1, \dots, X_p 和 Y 的样本相关系数矩阵

$$\mathbf{R} = \begin{bmatrix} r_{11} & \dots & r_{1p} r_{1y} \\ r_{21} & \dots & r_{2p} r_{2y} \\ \vdots & \dots & \vdots \\ r_{y1} & \dots & r_{yp} r_{yy} \end{bmatrix}, \text{ 同时}$$

$$r_{kj} = \frac{S_{kj}}{\sqrt{S_{kk}S_{jj}}} \quad \text{und} \quad r_{yk} = r_{ky} = \frac{S_{ky}}{\sqrt{S_{kk}S_{yy}}}$$

$$R = \sqrt{1 - r_{yy}}$$

$$R^2 = 1 - r_{yy}$$

$$R_{adj}^2 = R^2 - \frac{(1 - R^2)p}{C - p^*}$$

$$\Delta R^2 = R_{current}^2 - R_{previous}^2$$

$$\Delta F = \begin{cases} \frac{\Delta R^2(C - p^*)}{q(1 - R_{current}^2)} \\ \frac{\Delta R^2(C - p^* - q)}{q(R_{previous}^2 - 1)} \end{cases}$$

$$SS_e = r_{yy}(C - 1)S_{yy}$$

$$SS_R = R^2(C - 1)S_{yy}$$

$$AIC = C \ln \left(\frac{SS_e}{C} \right) + 2p^*$$

$$SBC = C \ln \left(\frac{SS_e}{C} \right) + p^* \ln(C)$$

$$VIF_i = \frac{1}{r_{ii}}$$

$$Tolerance_i = r_{ii}$$

$$b_k = \frac{r_{yk} \sqrt{S_{yy}}}{\sqrt{S_{kk}}} \text{ for } k = 1, \dots, p$$

$$b_k \pm \hat{\sigma}_{b_k} t_{0.975, C - p^*}$$

$$b_0 = \bar{y} - \sum_{k=1}^p b_k \bar{X}_k$$

$$Beta_k = r_{yk}$$

半偏相关

$$Part - Corr(X_k) = \frac{r_{yk}}{\sqrt{r_{kk}}}$$

偏相关

$$Partial - Corr(X_k) = \frac{r_{yk}}{\sqrt{r_{kk}r_{yy} - r_{yk}r_{ky}}}$$

杜宾 瓦森统计量

$$DW = \frac{\sum_{i=2}^l (\tilde{e}_i - \tilde{e}_{i-1})^2}{\sum_{i=1}^l c_i \tilde{e}_i^2}, \text{ 同时 } \tilde{e}_i = e_i \sqrt{g_i}.$$

寿命表（保险精算法）公式

SPSS 过程 SURVIVAL

计算方法

 X_j 从开始事件到目标事件或事件 j 截尾的时间 w_j 案例 j 的权重 K

区间数量

 t_i 第 i 个区间开始的时间 h_i 区间 i 的宽度 c_i 在区间 i 中截尾的案例的加权和 d_i 在区间 i 中发生目标事件的案例的加权和

计算区间，对事件和截尾得分

$$t_i \leq X_j < t_{i+1}$$

输出量数量（“生存”）

$$l_i = l_{i-1} - c_{i-1} - d_{i-1}$$

受到一个目标事件风险的数量

$$r_i = l_i - c_i / 2$$

失效（“死亡”）元素的比例

$$q_i = \frac{d_i}{r_i}$$

剩余（“生存”）元素的比例

$$p_i = 1 - q_i$$

区间末端剩余元素的累积比例

$$P_i = P_{i-1} p_i$$

中位生存时间

若 $P_k > 0.5$ ，则 $t_k +$ ，否则

$$Md = (t_i) + \frac{h_{i-1}(P_{i-1} - 0.5)}{P_{i-1} - P_i}$$

参考文献

- Allison, Paul D. (2001⁵). *Survival Analysis using the SAS System*. Cary, NC: SAS Institute Inc.
- Altman, Douglas G. (1992). *Practical Statistics for Medical Research*. London: Chapman & Hall/CRC.
- Ayres, Ian (2007). *Super Crunchers: How anything can be predicted*. London: John Murray Publ.
- Barnett, Jane (2008). Link between online gaming and violence killed off. Paper presented at the British Psychological Society's Annual Conference (Dublin, 02.04.2008).
- Berry, Michael J.A. & Linoff, Gordon, S. (2000). *Mastering Data Mining: The Art and Science of Custer Relationship Management*. New York: John Wiley & Sons.
- Bland, Martin (1995³). *An Introduction to Medical Statistics*. Oxford: Oxford University Press.
- Boehmke, Frederick J.; Morey, Daniel S. & Shannon, Megan (2006). Selection bias and continuous-time duration models: Consequences and a proposed solution. *American Journal of Political Science* 50, 1, 192-207.
- Böhning, Dankmar (1998). *Allgemeine Epidemiologie und ihre methodischen Grundlagen*. München Wien: R.Oldenbourg Verlag.
- Borg, Walter R. & Gall, Meredith D. (1989⁵). *Educational Research: An Introduction*. White Plains, NY: Longman.
- Bortz, Jürgen (1993⁴). *Statistik für Sozialwissenschaftler*. Heidelberg: Springer.
- Box-Steffensmeier, Janet M. & Jones, Bradford S. (1997). Time is of the essence: Event history models in political science. *American Journal of Political Science*, 41, 4, 1414-1461.
- Box-Steffensmeier, Janet M. & Jones, Bradford S. (2004). *Event history modeling: A guide for social scientists*. NY: Cambridge University Press.
- Bredenkamp, Jürgen (1972). *Der Signifikanztest in der psychologischen Forschung*. Frankfurt/M.:

Akademische Verlagsanstalt.

Cha, Kwang Y.; Wirth, Daniel P.; Lobo, Rogerio A. (2001). Does prayer influence the success of in vitro fertilization-embryo transfer? *Journal of Reproductive Medicine*, 46, 781-787.

Chapman, Pete; Clinton, Julian; Khabaza, Thomas; Reinartz, Thomas; Wirth, Rüdiger (1999). The CRISP-DM Process Model. Discussion Paper. The CRISP-DM consortium NCR System Engineering Copenhagen (Denmark), DaimlerChrysler AG (Germany), Integral Solutions Ltd. (England) and OHRA Verzekeringen en Bank Groep B.V (The Netherlands).

Chatterjee, Samprit & Price, Bertram (1995). *Praxis der Regressionsanalyse*. München Wien: R.Oldenbourg Verlag.

Cohen, Jacob et al. (2003³). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah NJ: Lawrence Erlbaum Ass.

Collett, David (2003²). *Modeling Survival Data in Medical Research*. London: Chapman & Hall/CRC.

Cox, David R. (1972) Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, 34, 187-220.

Cox, David R. & Snell, Joyce E. (1989). *The Analysis of Binary Data*. London: Chapman and Hall.

Cox, David R. & Oakes, David (1984). *Analysis of Survival Data*. London: Chapman & Hall/CRC.

Darlington, Richard B. (1990). *Regression and Linear models*. New York: McGraw-Hill.

Diehl, Joerg & Kohr, Heinz (1999). *Deskriptive Statistik*. Eschborn bei Frankfurt/M.: Verlag Dietmar Klotz.

Elandt-Johnson, Regina C. & Johnson, Norman L. (1999). *Survival Models and Data Analysis*. New York: John Wiley & Sons.

Eliason, Scott R. (1993). *Maximum-Likelihood Estimation: Logic and Practise*. (Series: Quantitative Applications in the Social Sciences). Thousand Oaks: Sage Publications.

Elmore, Patricia B. & Woehlke, Paula L. (1998). Research Methods employed in “American Educational Research Journal”, “Educational Researcher”, and “Review of Educational Research” from 1978 to 1995 (Paper presented at the Annual Meeting of the American Educational Research Association, San Diego (CA), April 13-17, 1998).

Elmore, Patricia B. & Woehlke, Paula L. (1996). Research Methods employed in “American Educational Research Journal”, “Educational Researcher”, and “Review of Educational Research” from 1978 to 1995 (Paper presented at the Annual Meeting of the American Educational Research Association, New York (NY), April 8-12, 1996).

Ferguson, Christopher J. (2007). Evidence for publication bias in video game violence effects literature: A meta-analytic review. *Aggression and Violent Behavior*, 12, 4, 470-482.

- Finney, David J. (1996). A note on the history of regression. *Journal of Applied Statistics*, 23, 5, 555-558.
- Gale, Catharine R.; Deary, Ian J.; Schoon, Ingrid & Batty, G. David (2007). IQ in childhood and vegetarianism in adulthood: 1970 British cohort study. *British Medical Journal*, February 3, 334 (7587), 245.
- Goertzel, Ted (2002). Myths of murder and multiple regression. *The Skeptical Inquirer*, 26, 1/2, 19-23.
- Goodwin, Laura D. & Goodwin, William L. (1985). An analysis of statistical techniques used in the *Journal of Educational Psychology*, 1979-1983. *Educational Psychologist*, Volume 20, 1, 13-21.
- Graber, Marion (2000). Data Mining: Eine mächtige Methode im Business-Intelligence-Prozess. *IT-Management*, 1/2, 1-6 (Sonderdruck).
- Green, Sam (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research*, 26, 455-51.
- Guggenmoos-Holzmann, Irene & Wernecke, Klaus-Dieter (1996). *Medizinische Statistik*. Berlin: Blackwell Wissenschaftsverlag.
- Hackbarth, Diana (2008). Research Reporting and Evidence of Effectiveness: Why "No Difference" Matters. *American Journal of Critical Care*, 17(3), 218-220.
- Hartung, Joachim (1999¹²). *Statistik*. München Wien: R. Oldenbourg Verlag.
- Hess, Kenneth R. (2007). Graphical methods for assessing violations of the proportional hazards assumption in cox regression. *Statistics in Medicine*, 14, 15, 1707-1723.
- Hoerl, Arthur E. & Kennard, Robert W. (1970). Ridge-Regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 69-82.
- Hosmer, David W. & Lemeshow, Stanley (2000). *Applied Logistic Regression*. Second Edition. Wiley & Sons: New York.
- Hosmer, David W. & Lemeshow, Stanley (1999). *Applied Survival Analysis*. John Wiley and Sons: New York.
- Howarth, Richard J. (2001). A history of regression and related model-fitting in the earth sciences (1636?-2000). *Natural Resources Research*, 10, 4 (12), 241-286.
- Hsu, Tse-chi (2005). Research methods and data analysis procedures used by educational researchers. *International Journal of Research & Method in Education*, 28, 2, 109-133.
- Hulland, John (1999). Use of Partial Least Squares (PLS) in Strategic Management Research: A review of four recent studies. *Strategic Management Journal*, 20, 195-204.
- Jaccard, James (2001). *Interaction Effects in Logistic Regression Analysis (Series: Quantitative Applications in the Social Sciences)*. Thousand Oaks: Sage Publications.

- Kahn, Harold A. & Sempos, Christopher T. (1989). *Statistical Methods in Epidemiology*. New York - Oxford: Oxford University Press.
- Kalbfleisch, John D. & Prentice, Ross L. (2002). *Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons.
- Kaplan, Edward L. & Meier, Paul (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457-481.
- Khabaza, Tom (2005). *Hard hats for data miners: Myths and pitfalls of data mining*. Chi-cago: SPSS Inc.
- Klecka, William R. (1980). *Discriminant Analysis. Quantitative Applications in the Social Sciences Series, No. 19*. Thousand Oaks, CA: Sage Publications.
- Klein, John P. & Moeschberger, Melvin L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag: New York.
- Kleinbaum, David G. & Klein, Mitchel (2005²). *Survival Analysis: A Self-Learning Text*. Springer-Verlag: New York.
- Kutner, Lawrence & Olson, Cheryl K. (2008). *Grand Theft Childhood: The surprising truth about violent video games and what parents can do*. New York: Simon & Schuster.
- Lawless, Jerald F. (2002²). *Statistical Models and Methods for Lifetime Data*. New York: John Wiley & Sons.
- Lee, Elisa T. (1992²). *Statistical Methods for Survival Data Analysis*. New York: John Wiley & Sons.
- Litz, Hans Peter (2000). *Multivariate statistische Methoden*. München: Oldenburg.
- Lorenz, Rolf J. (1992³). *Grundbegriffe der Biometrie*. Stuttgart: Gustav Fischer Verlag.
- Mantel, Nathan (1970). Why stepdown procedures in variable selection. *Technometrics*, 12, 3, 621-625.
- Menard, Scott (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54, 17-24.
- Menard, Scott (2001²). *Applied Logistic Regression Analysis (Series: Quantitative Applications in the Social Sciences)*. Thousand Oaks: Sage Publications.
- McFadden, Daniel (2004). *Persönliche Information*, 27.01.2004.
- Nagelkerke, Nico J. D. (1991). A note on the general definition of the coefficient of determination. *Biometrika*, 78, 691-692.
- Nelson, Wayne (1981). Analysis of Performance-Degradation Data. *IEEE Transactions on Reliability*. Vol. 2, R-30, No. 2, 149-155.

- Pearson, Karl (1896). Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. *Philosophical Transactions of the Royal Society of London*, 187, 253-318.
- Pedhazur, Elazar J. (1982²). *Multiple Regression in Behavioral Research: Explanation and Prediction*. Fort Worth: Holt, Rinehart and Winston Inc.
- Pötschke, Manuela & Simonson, Julia (2003). Konträr und ungenügend? Ansprüche an Inhalt und Qualität einer sozialwissenschaftlichen Methodenausbildung. *ZA-Information*, 52, 72-92.
- Press, S. James & Wilson, Sandra (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, Vol. 73: 699-705.
- Rasch, Dieter; Herrendörfer, Günter; Bock, Jürgen; Victor, Norbert und Guiard, Volker (Hrsg.) (1996). *Verfahrensbibliothek: Versuchsplanung und -auswertung. Band I*. München Wien: R.Oldenbourg Verlag.
- Rasch, Dieter, Herrendörfer, Günter, Bock, Jürgen, Victor, Norbert und Guiard, Volker (Hsg.) (1998). *Verfahrensbibliothek: Versuchsplanung und -auswertung. Band II*. München Wien: R.Oldenbourg Verlag.
- Raubenheimer, Jenny E. (2004). An item selection procedure to maximize scale reliability and validity. *South African Journal of Industrial Psychology*, 30, 4, 59-64.
- Rexer, Karl; Gearan, Paul & Allen, Heather N. (2007). *Surveying the Field: Current Data Mining Applications, Analytic Tools, and Practical Challenges*, Data Miner Survey Summary Report, August 2007. Source: www.RexerAnalytics.com.
- Rigby, Alan; Armstrong, Gillian; Campbell, Michael & Summerton, Nick (2004). A survey of statistics in three UK general practice journals. *BMC Medical Research Methodology*, 13, 28.
- Ripoll, Ramón Mora; Terren, Carlos Ascaso; Vilalta, Joan Sentís (1996). The current use of statistics in biomedical investigation: A comparison of general medicine journals. *Medicina Clinica*, 106, 451-456.
- Rud, Olivia (2001). *Data Mining Cookbook*. New York: John Wiley & Sons.
- Ryan, Thomas P. & Woodall, William H. (2005). The Most-Cited Statistical Papers. *Journal of Applied Statistics*, 32, 5, 461-474.
- Sarris, Viktor (1992). *Methodologische Grundlagen der Experimentalpsychologie. Band 2: Versuchsplanung und Stadien des psychologischen Experiments*. München: UTB Reinhardt.
- Schendera, Christian FG (2007). *Datenqualität mit SPSS*. München Wien: R.Oldenbourg Verlag.
- Schendera, Christian FG (2005). *Datenmanagement mit SPSS*. Heidelberg: Springer.
- Schendera, Christian FG (2004). *Datenmanagement und Datenanalyse mit dem SAS System*. München: Oldenburg.

Schlittgen, Rainer & Streitberg, Bernd H.J. (2001⁹). *Zeitreihenanalyse*. München Wien: R.Oldenbourg Verlag.

Schlittgen, Rainer (2001). *Angewandte Zeitreihenanalyse*. München Wien: R.Oldenbourg Verlag.

SPSS (2007a). *SPSS 16.0 Command Syntax Reference*. Chicago: SPSS Inc.

SPSS (2007b). *Clementine 12.0.Modeling Nodes*. Chicago: SPSS Inc.

SPSS (2007c). *SPSS 16.0 Algorithms*. Chicago: SPSS Inc.

Stanton, Jeffrey M. (2001). Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors. *Journal of Statistics Education*, Vol. 9, No. 3 [online].

Therneau Terry M. & Grambsch, Patricia M. (2002²). *Modeling Survival Data: Extending the Cox Model*. Heidelberg: Springer.

Tobias, Randall D. (1997). *An introduction to partial least squares regression*. Cary, NC: SAS Institute.

Turner, Erick H. ; Matthews, Annette M. ; Linardatos, Eftihia ; Tell, Robert A. & Rosenthal, Robert (2008). Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy. *The New England Journal of Medicine*, January, 3, Vol. 358:252-260.

Tutz, Gerhard (2000). *Die Analyse kategorialer Daten*. München Wien: R.Oldenbourg Verlag.

Vinzi, Vincenzo E.; Chin, Wynne W.; Henseler, Joerg & Wang, Huiwen (2008) (Eds.). *Handbook of Partial Least Squares: Concepts, methods and applications in marketing and related fields*. Series: Springer Handbooks of Computational Statistics. New York: Springer-Verlag.

Weinzimmer, Laurence G.; Mone, Mark A. & Alwan, Layth C. (1994). An examination of perceptions and usage of regression diagnostics in organization studies. *Journal of Management*, 20, 1, 179-192.

Wentzell, Peter D. & Vega Montoto, Lorenzo (2003). Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures. *Chemometrics and Intelligent Laboratory Systems*, 65, 2, 257-279.

West, Prof. Stephen G. (New York, pers. Kommunikation 05.09.2006).

Witte, Erich H. (1980). *Signifikanztest und statistische Inferenz: Analysen, Probleme, Alternativen*. Stuttgart: Enke.

Wold, Herman (1981). *The fix-point approach to interdependent systems*. Amsterdam: North Holland.

Wold, Herman (1985). Partial Least Squares, 581-591; in: Kotz, Samuel & Johnson, Norman L. (Eds.), *Encyclopedia of Statistical Sciences*, Vol. 6, New York: Wiley.

Wold, Herman (1994). PLS for multivariate linear modeling. In: Van der Waterbeemd, H. (Ed.). *QSAR: Chemometric methods in molecular design: Methods and principles in medicinal chemistry*. Weinheim: Verlag-Chemie.

Woolridge, Jeffrey M. (2003²). *Introductory econometrics: A modern approach*. Mason, Ohio: South Western.

Yaffee, Robert & McGee, Monnie (2000). *Times Series Analysis and Forecasting*. Orlando/Fl.: Academic Press.

您对本书的建议和意见

著书之时，笔者力求使本书内容全面、易懂、精准且不失其时效性。即使做了多次检查、校对，可能或多或少还会存在不准确或者不易理解之处，只能在后续版本中解决这些问题。而且 SPSS 软件的技术和统计分析方法也在不断发展，将来也会予以考虑。

在此，笔者衷心希望各位读者能够就书中的不足之处提出改进的建议和意见，并发送邮件至 SPSS3@method-consult.de。

请您务必在邮件主题中注明：“对 SPSS 一书的回馈建议”，并标出版本、页码、关键词（例如：错别字）以及对问题的描述（例如，在统计分析中）。若为编程代码问题，请您附上批注。

衷心感谢！

克里斯蒂安•FG•申德拉博士

作者简介

要对事物进行认知与了解，采取方法是非常重要的。而只有明确了相关的研究方法，才能对认知和了解做出评估，并对建立在此基础上的决策的后果及其水平进行估计。

作者简介

克里斯蒂安•FG•申德拉博士（Dr. Christian FG Schendera）先生致力于对知识的理性构建与重建，即各种科学的和不科学的研究方法（尤其是统计方法）对人们构建知识和接受知识有什么样的影响。

申德拉先生目前担任瑞士计算机科学公司（CSC）商业智能团队的科学数据分析师和主管。从业领域主要包括高级分析（数据分析/数据挖掘）、科学咨询（科学方法咨询）以及 SPSS 和 SAS 培训。他的客户大多是来自不同行业（例如，银行与保险、市场营销及医药）、不同国家（主要是德国、奥地利和瑞士）的企业。曾经参与了多个科研、分析和评估项目。出版过数本数据分析、数据质量、SAS 和 SPSS 的相关著作。例如，关于 SPSS 培训的详情请参见 www.ch.csc.com。

卢塞恩师范大学简介

卢塞恩师范大学是瑞士最大的师范类高校之一，在规模方面有非常丰富的专业和课程设置，但另一方面又比较小，可以让人与人之间建立紧密的联系。目前，卢塞恩师范大学有大约 1600 名注册大学生。此外，每年有 5500 名教师来这里参加丰富多彩的培训项目。卢塞恩师范大学对于教师的培训既贴近实践又有深厚的学术基础，并为他们今后的职业发展提供帮助。此外，该大学还通过专项的科研工作支持在教育领域的知识产生和知识交流，并为瑞士中部地区的教育行业提供各种服务。学校的宗旨是既满足实践的需求，又紧跟科学的最新发展，并且与国内外的很多高校、专科院校和大学建立了合作关系。卢塞恩师范大学的财政支出由卢塞恩州政府承担。

详细信息请参阅 www.phlu.ch。